JOINT MODELLING OF SURVIVAL AND LONGITUDINAL OUTCOMES OF HIV/AIDS PATIENTS IN LIMPOPO PROVINCE, SOUTH AFRICA

by

KHEHLA DANIEL MOLOI

THESIS

Submitted in fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: PROF. YEHENEW GETACHEW KIFLE

CO-SUPERVISOR: PROF. KHANGELANI ZUMA (HSRC, SOUTH AFRICA)

2019

Declaration

I, **Khehla Daniel Moloi**, declare that the thesis hereby submitted to the University of Limpopo, for the degree of Doctor of Philosophy (PhD) in Statistics has not been submitted by me before or anyone for a degree at this or any other University. This is my original work in design and execution, and that all material contained herein has been duly acknowledged throughout the thesis.

Signature:........................... Date:................................

Khehla Daniel Moloi

Dedication

This PhD work is dedicated to:

My late mother and father:

Evelyn and Daniel

My wife:

Manoko F Moloi

My children:

Moratuwa Nthabiseng,
Lebohang Boitumelo (late), and
Lehasa Polane Moloi

Acknowledgments

I would like to give thanks, praise to Almighty God for the strength, patience, wisdom and courage He instilled in me throughout the period of undertaking this study. Also, I would like to thank the Holy Spirit in leading and guiding me in times of hardship.

I would wish to express my sincere gratitude to my supervisor Professor Yehenew Gatechaw Kifle (University of Limpopo, South Africa) for his patience, professional guidance, and academic support throughout the writing of this thesis. He consistently allowed me to own this thesis, but steered me in the right direction whenever he thought I need it. My co-supervisor Professor Khangelani Zuma (HSRC, South Africa) for his support and guidance.

I am also thanful to the Limpopo Department of Health in Limpopo Province for allowing me to use their HIV/AIDS dataset. Special gratitude to Dr. E. R. Andrinopoulou (University of Erasmus Medical Centre, Rotterdam, Netherlands) for her guidance and contribution in writing this thesis. Many thanks to NRF-TDG reference number APP-TDG-071 for funding my visit to University of Erasmus Medical Centre, Rotterdam, Netherlands, for collaborative research. The University of Limpopo is also acknowledged granting me with this opportunity to undertake this study. The Faculty of Science and Agriculture and Department of Statistics and Operations Research for funding my visit to University of Maryland Baltimore County, United States of America. Professor DoHwan Park and Professor Bimal Sinha from University of Maryland Balti-

more County, United States of America, as readers of this thesis, and I am deeply indebted to their valuable comments. Mr. Barry Mutasa from ANOVA Health Institute: Tzaneen, Limpopo Province, South Africa for his valuable contribution.

Last but by no means the least special appreciation to my beloved family and especially spouse for providing unfailing support and continuous encouragement throughout my years of study and through the process and writing this thesis. m.

Thank You All!

Contents

List of Figures

List of Tables

Acronyms

AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
ART	Antiretroviral Therapy
ARV	Antiretroviral
HAART	Highly Active Antiretroviral Therapy
HIV	Annual Human Immunodeficiency Virus
GBV	Gender Base Violence
HR	Hazard Ratio
IGC	Individual Growth Curve
KM	Kaplan-Meier
LGBTI	Lesbian, Gay, Bisexual, Transgender and Intersex
ML	Maximum Likelihood
LME	Linear Mixed Effects
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
REML	Restricted Maximum Likelihood
PH	Proportional Hazard
PMTCT	Prevention of Mother to Child Transmissions
SANAC	South African National AIDS Council
UNAIDS	Joint United Nations Programme on HIV/AIDS
WHO	World Health Organisation

Chapter 1

General Introduction

1.1 Motivating Case Studies

1.1.1 Definition of HIV/AIDS

Human immunodeficiency virus (HIV) is a virus that attacks the immune system, which is our body's natural defence against illness. The virus destroys a type of white blood cells in the immune system called a T-helper cell, and reproduce itself inside these cells. The T-helper cells are also referred to as CD4 cells. As HIV themselves destroys more CD4 cells and multiply themselves, it gradually breaks down a person's immune system. If HIV is left untreated, it may take 10 to 15 years for the immune system to be so severely damaged to the extend that it can no longer defend itself at all. The speed of HIV progresses depends on age, health and background of a patient [1].

Acquired Immune Deficiency Syndrome (AIDS) is caused HIV. AIDS stand for: *Acquired* means one gets infected with it; *Immune Deficiency* means a weakness in the body's system that fights diseases; *Syndrome* means group of health problems that makes up a diseases. A person is said to have AIDS when their immune system is too weak to fight off infection, and they develop certain defining symptoms and illnesses. This is the last clinical stage of HIV, when infection is advanced, and if left untreated will lead to death.

1.1.2 Transmission and Prevention of HIV/AIDS

HIV is transmitted from an infected person to another through direct contact of bodily fluids such as:

i) blood transfusion; pre-seminal fluids; rectal fluids; vaginal fluids;

ii) sharing a needle or syringe with an infected person since HIV can live in a used needle for 42 days depending on the temperature and other factors;

iii) a baby being born when their mother is infected or drinking milk of infected mother [2].

A HIV negative persons can avoid being infected by:

i) avoiding the use of recreational drugs which share needle or syringes;

ii) reacting quickly if you believe that you have become exposed, through the use of post-exposure prophylaxis (PrEP) treatment. If treated within 72 hours you may be able to prevent HIV infection;

iii) avoiding contact with other people's blood and certain other bodily fluids that can spread HIV [3].

Mother-to-child transmission of HIV is the most common way young babies contract the virus and happens when HIV is passed from a mother to her unborn baby during pregnancy, birth or breastfeeding. As a result, the South African National Department of Health came up with an effective prevention-of- mother-to-child-transmissions (PMTCT) programme that requires mothers and their babies to:

i) receive antenatal services and HIV testing during pregnancy;

ii) have access to antiretroviral treatment (ART);

iii) practice safe childbirth practices and appropriate infant feeding;

iv) make use of infant HIV testing and other post-natal health care services.

South Africa's health guidelines for HIV/AIDS divided the national programme into the following three broad categories, namely Category 1: Antenatal care during pregnancy, Category 2: Labour and delivery and Category 3: Postnatal care after delivery.

Category 1 (Antenatal care during pregnancy): When a pregnant woman visits a clinic, a midwife asks her to take a voluntary HIV testing. If she is found to be HIV positive, she is offered routine HIV counselling, and she has an option of joining PMTCT programme for free. All HIV positive pregnant women or breastfeeding qualify for ART, and the ART'S reduces the risk of mother to baby transmission and protects the mother's health during and after pregnancy. The ART'S should be taken as soon as possible after diagnosis preferably within seven days [12].

Category 2 (Labour and delivery): When an HIV positive woman continues to take ART'S throughout her pregnancy, she is able to deliver natural, and, as a result transmission of HIV from mother to child is prevented.

Category 3 (Postnatal care after delivery): Women continue to take their treatment as normal. All babies are given a dose of ART's daily up until the mother stop breastfeeding. The baby is also given antibiotic treatment to prevent infection. Breastfeeding is generally encouraged, however, mothers are given a choice to either breastfeed for six months or give their babies formula feed [5].

1.1.3 Status of HIV/AIDS in South Africa

HIV/AIDS is a major concern throughout the world. Though Sub-Saharan Africa is the most affected region. However, no country in the world can ignore the problem. This PhD study also focused on applying various modelling techniques to the HIV/AIDS dataset collected from a province located in the northern part of South Africa, namely Limpopo (Figure 1.1).

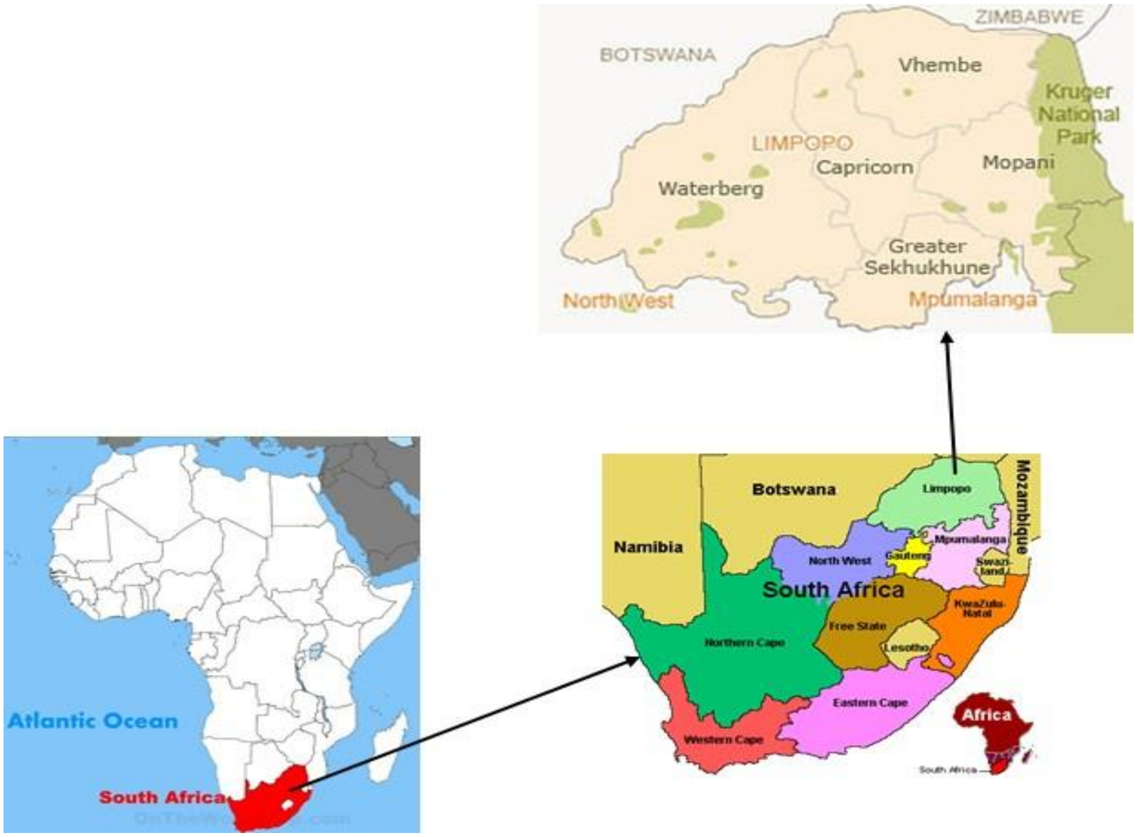


Figure 1.1: Study area Map.

Figure 1.2 below shows upwards trend of HIV prevalance from 2002 to 2015, indicating that the spread of HIV/AIDS was increasing each year in South African population, whereas, the incidence was decreasing tremendously. The continual decrease in HIV incidence and continual increase in HIV prevalence can be attributed to the fraction of individuals that are affected remains high. Table 1.1 shows that South Africa has the biggest and highest profile of HIV epidemic in the whole world, with an estimated 6.2 million people living with HIV in 2015. In the same year, there were approximately 380,000 new infections while 180,000 died from AIDS-related illnesses [12].

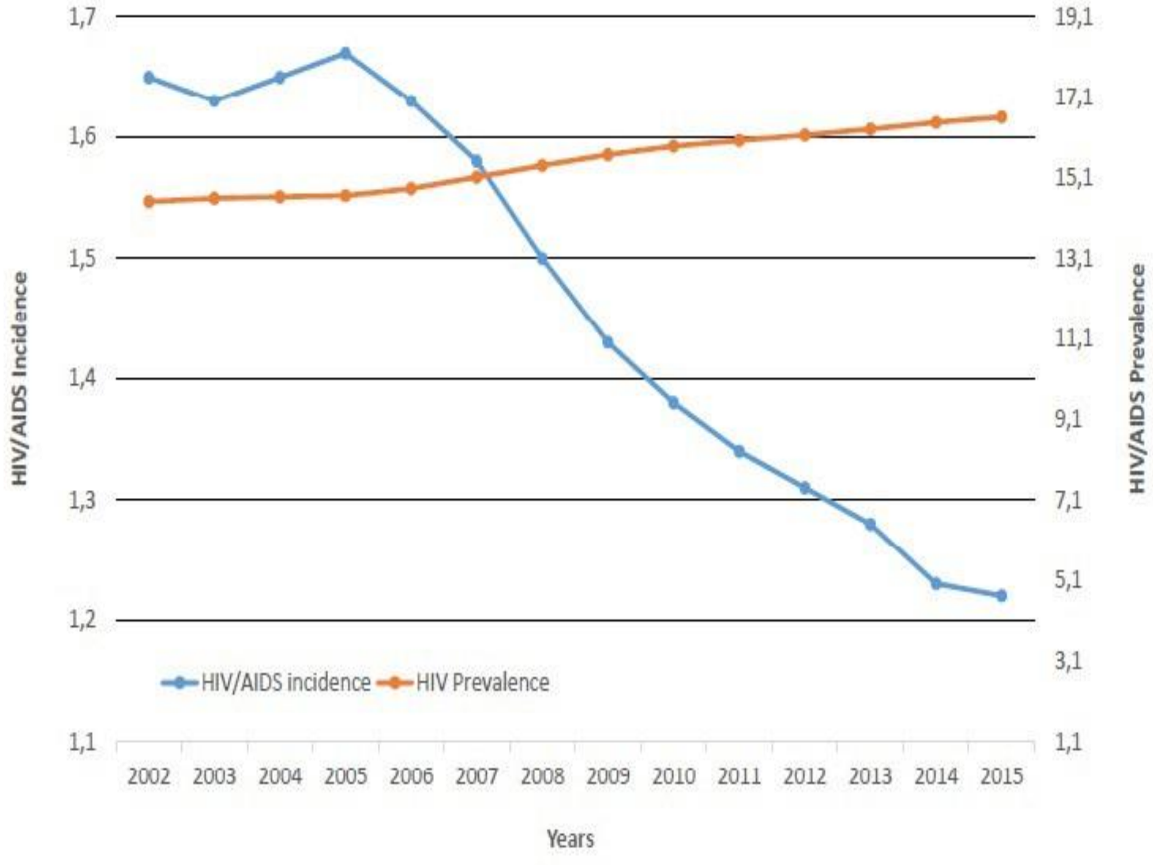


Figure 1.2: Trend of HIV/AIDS incidence and prevalence from 2002 to 2015 in South Africa.

Table 1.1 shows the HIV prevalence estimates and the total number of people living with HIV from 2002 to 2015. The total number of people living with HIV/AIDS in South Africa increased from an estimated 4.02 million in 2002 to 6.19 million in 2015. In 2015, it was estimated that 11.22% of the total population was HIV positive.

Table 1.1: HIV/AIDS Prevalence, incidence and number of people living with HIV/AIDS in South Africa from 2002 to 2015.

Year	Prevalence (in %)			Incidence (in %)	Population with HIV/AIDS (in millions)	Total Pop. (in %)
	Women	Adults	Youth			
2002	16.69	14.5	6.75	1.65	4.02	8.8
2003	16.85	14.58	6.35	1.63	4.14	9.0
2004	16.93	14.62	6.07	1.65	4.25	9.1
2005	17.01	14.65	5.91	1.67	4.35	9.2
2006	17.22	14.82	5.82	1.65	4.51	9.4
2007	17.52	15.1	5.76	1.58	4.71	9.7
2008	17.81	15.39	5.71	1.5	4.93	10.0
2009	18.09	15.66	5.69	1.43	5.13	10.2
2010	18.29	15.87	5.7	1.38	5.32	10.4
2011	18.42	16.01	5.64	1.34	5.48	10.6
2012	18.43	16.14	5.61	1.31	5.65	10.7
2013	18.67	16.29	5.6	1.28	5.83	10.9
2014	18.85	16.46	5.59	1.23	6.02	11.1
2015	18.99	16.59	5.59	1.22	6.19	11.2

Source: http://www.tbfacts.org/hiv-statistics-south-africa

However, South Africa has made huge improvement in getting people to test for HIV in recent years and is now almost meeting the first of the 90-90-90 targets with 86% of people aware of their HIV status. This is inline with the 2014 UNAIDS global target that 90% of all HIV-positive persons be aware of their HIV status, provide antiretroviral therapy (ART) for 90% of those diagnosed, and achieve viral suppression for 90% of those treated by 2020 [11].

The groups that are most affected by HIV in South Africa are:
Group 1 (Sex workers): The national HIV prevalence among sex workers is estimated at 57.7%. The study done by Slabbert *et al.*, (2017) found that poverty, number of dependents they have and lack of alternative career opportunities were the cause for increase HIV risk for South Africa sex workers. The use of injection also exacerbated their vulnerability to HIV infection.

Group 2 (Men who have sex with men): HIV prevalence among men who have sex with men in South Africa is now estimated at 26.8% [1]. Despite the existence of South African constitution that protects the rights of lesbian, gay, bisexual and transgender communities, many men who have sex with men faces high level of social stigma and homophobic violence as result of traditional and conservative attitudes within the general population [8].

Group 3 (Transgender women): The transgender women in Sub-Sahara Africa are twice as likely to have HIV as men who have sex with men [9]. This community has been neglected by both the policy and research in South Africa, either they have been excluded on the bases mis-categorisation as men who have sex with men. However, the South African AIDS Council's lesbian, gay, bisexual, transgender, and intersex (LGBTI) and HIV Framework recognise transgender women as a key affected population.

Group 4 (People who inject drugs): A 2016 study of people who inject drugs in five South Africa cities found 32% of men and 26% of women regularly share syringes and other injection equipment and nearly half those re-use needles [10].

Group 5 (Children and orphans): In 2016 study found that, approximately 320,000 children (aged 0 to 14) were living with HIV in South Africa, only 55% of those were on treatment. However, new infections have declined among South African children, from 25,000 in 2010 to 12,000 in 2016. This is mainly due to the success of PMTCT programmes. The rate of mother-to-child transmission stood at 1.3% in 2017, down from 3.6% in 2011. This indeed puts South Africa on track for eliminating mother to child transmission of HIV [11].

Group 6 (Women, adolescent girls): HIV prevalence among young women in South Africa is nearly four times greater than that of men their age. Young women between the ages of 15 and 24 made up 37% of new infections in South Africa in 2016. To try and reduce this high rate of infection, young women and adolescent girls who are considered at high risk of HIV infection are now being offered pre-exposure prophylaxis (PrEP) [12]. In may research literature, poverty, the low status of women and gender-based violence (GBV) have all been cited as reasons for the disparity in HIV prevalence between gender. Indeed GBV attributable to an estimated 20–25% of new HIV infections in young women.

South Africa has the largest antiretroviral treatment (ART) programme globally and these programmes has been largely funded from its domestic resources. In 2015, South Africa has invested more than 19,5 billion Rand annually to run its HIV/AIDS programmes [11]. Table 1.2 below shows that there approximately 1,793,000 HIV positive persons in South Africa who were on ART programme in 2011, of which, 1,525,000 were on treatment in the public sector, and more than 190,000 were on treatment in the private sector. While Non-Governmental Organisation (NGOs) had more than 78,000 HIV persons on their ART programme.

Despite all these efforts from the National Department of Health, private sector and NGOs, HIV prevalence still remains high (19.2 %) among the general population, although it varies markedly between provinces. There is a substantial difference in HIV among the nine provinces and HIV statistics shows that the difference has been consistent over a number of years [13].

In Table 1.3, statistics for South Africa shows that there continue to be a very high number of HIV related deaths. In 2015 the number of people who died from HIV related illnesses was 162,445 which was 30.5% of all deaths in South Africa.

Table 1.2: HIV/AIDS patients on ART in South Africa by gender and service provider from 2007 to 2011.

		Year				
		2007	2008	2009	2010	2011
Gender	Men	120,000	183,000	283,000	396,000	551,000
	Women	228,000	354,000	553,000	777,000	1,090,000
	Children	34,000	51,000	76,000	114,000	152,000
	Total	382,000	588,000	912,000	1,287,000	1,793,000
Provider	Public Sector	290,000	470,000	748,000	1,073,000	1,525,000
	Private Sector	68,000	86,000	117,000	154,000	190,000
	NGOs	24,000	32,000	47,000	60,000	78,000
	Total	382,000	588,000	912,000	1,287,000	1,793,000

Source:http://www.tbfacts.org/hiv-statistics-south-africa

Table 1.3: Deaths from HIV/AIDS related illnesses in South Africa 2010-2015.

Year	HIV/AIDS related deaths	Total Number of deaths	Deaths that are from HIV/AIDS related illness (in %)
2010	183 465	535 396	34.3
2011	200 654	556 087	36.1
2012	197 090	555 921	35.5
2013	177 624	539 880	32.9
2014	151 040	516 929	29.2
2015	162 445	531 965	30.5

Source:http://www.tbfacts.org/hiv-statistics-south-africa

1.1.3.1 Factors contributing to the spread of HIV

The main factors that contribute to the spread of HIV are the followings: poverty; inequality and social instability; high level of sexually transmitted infections; the low status of women; sexual violence; high mobility (particularly migrant labour); limited and uneven access to quality medical care; a history of poor leadership in the response to the epidemic and society leaders dying and leaving a generation of children growing up without the care and role models they will normally have. These afore-mentioned factors seem to be applicable even here Limpopo Province, for example, in Phalaborwa town, which is a platinum mine town, HIV prevalence is very high because of migrant labourers. In addition to above mentioned factors, many people in Limpopo Province do not know their HIV status because its discussion is often a taboo. Educating the society of how a person is infected and how HIV is transmitted will play a vital role in reducing the spread of HIV positive persons. Again, testing and counselling the infected person will help in reducing spread of HIV [14].

1.2 Problem Statement

In most HIV/AIDS and related follow-up studies in South Africa and elsewhere in the world, it is common to collect data on repeated biomarker measurements, such as CD4 counts, viral load, as well as time-to-event variables, like time to death and time to default [15]. Moreover, such longitudinal HIV/AIDS data are clustered at various levels, including hospitals where patients are repeatedly followed-up [16]. Even though joint modelling is a complex modelling approach to handle such situations, it enables both longitudinal repeated biomarker measurements and survival possesses to be modelled together while taking into account the association between them. That is, by including a random effect for longitudinal data in the survival model, the patterns of a biomarker's performance and relationship between its progression and survival time can be characterised in a better way. In this way, joint modelling provides less biased estimates and more efficient inferences than separate models.

Many data analysts or medical researchers analyse survival and longitudinal data separately, which results in more biased estimates for their parameters. Moreover, in the presence of clustering in the data, such as the one in the South African HIV/AIDS dataset, this problem becomes even worse. In the literature, depending on the structure of the data at hand, there are a number of joint modelling approaches to be applied. These includes: shared parameter joint models, latent class joint models, and joint models in the presence of clustering. To our knowledge there is significantly less or little research work done in the South African context using various joint modelling approaches, that can handle different structures in the HIV/AIDS dataset with the aim of coming up with a good model that will improve the prediction of time-to-event variable, by considering all markers in the data and also enable the researcher to assess the joint evolution of the process at large.

1.2.1 Purpose of the study

The purpose of this study is to apply various modelling techniques to HIV/AIDS data in Limpopo Province, South Africa. Time to progression to AIDS or death is also recorded for each patient, although some subjects may withdraw early from the study or fail to experience the event by the time of study closure. In Chapter 2, analyses of HIV/AIDS data of Limpopo Province patients using survival data analyses is the main focus. In survival analyses we considered semi-parametric Cox model as well as parametric models and compare them. In Chapter 3 focused on analyses of HIV/AIDS longitudinal data (viral load) of Limpopo Province using linear mixed-effects models. A transformed logarithmic viral load will be a response variable, while our explanatory variables or covariates are: CD4 cell counts, age (at baseline), previous opportunistic infection (e.g., tuberculosis), gender, districts, health care facilities, and AIDS clinical stages.

In Chapter 4, focused on analyses HIV/AIDS data by using joint model of longitudinal and time-to-event data. Here our objective in longitudinal studies is to characterise the relationship between a longitudinal response (viral load) process and a time-to-event. Furthermore, basic joint models are explained, and it was assumed that the random effects underlie both longitudinal and survival process. Moreover joint latent class joint models were considered in this study as well as dynamic prediction of conditional probabilities for an event for any randomly selected patient in Limpopo Province based on 200 Monte Carlo samples. These various modelling approaches were compared using various statistical methods and thereby coming up with a reasonably good model that will handle both the survival and longitudinal processes simultaneously.

1.2.2 Aim

The aim of the study is to apply various joint modelling techniques to the clustered HIV/AIDS data in Limpopo Province, South Africa, in order to come up with a good model that will simultaneously handle the survival and longitudinal outcomes.

1.2.2.1 Objectives

The objectives of the study are:

i) to use Kaplan-Meier to compare the average evolutions between gender, districts, health care facilities, previous opportunistic infections, and AIDS clinical stages;

ii) to compare the semi-parametric and parametric models;

iii) to analysis survival data using both Cox proportional hazard and parametric hazard models;

iv) to describe the relationship between response variable and the covariates using linear mixed effect models;

v) to show how longitudinal evolution of viral load is associated with time-to-death;

vi) to characterise viral dynamics in patient population and intra- and inter-subject variation;

vii) to assume random effects that gives some structure to error terms that characterises individual variation due to some factor levels;

viii) to demonstrate non-linear statistical framework as a basis for estimation of population and individual viral dynamics parameters and how models may be used to draw biological relevant interpretations and aid clinical decision-making within the context of Limpopo Province HIV/AIDS data;

ix) perform separate longitudinal and survival analyses per outcome;

x) establish the strength of association between the longitudinal evolution of viral load and hazard rate to death;

xi) compare separate and joint models, and various association measures such as parametric joint and shared parameter models approach;

xii) Compare average evolutions between males and females;

xiii) show how marker-specific evolutions are related to each other (association of the evolution);

xiv) compute prediction for time to death for any randomly selected HIV positive patient by considering patient's viral load;

xv) come up with good joint model(s) that will handle simultaneously both the repeated measurements as well as the survival outcomes in the presence of clustering in the South African HIV/AIDS dataset; and

xvi) recommend to health decision-makers and policy-makers how the application of joint modelling techniques can be beneficiary to HIV/AIDS patients.

In Chapter 2, we will address the following objectives: Kaplan-Meier have been used as statistical tool to compare the average evolutions between gender, districts, health care facilities, previous opportunistic infections, and AIDS clinical stages;we will compare the semi-parametric and parametric models ; and, we will analyse survival data using both Cox proportional hazard and parametric hazard model; and finally compare semi-parametric and parametric models.

In Chapter 3 we will address the following objectives: we will describe the relationship between response variable and the covariates using linear mixed effect models and show how longitudinal evolution of viral load is associated with time-to-death. We will also characterise viral dynamics in patient population and intra- and inter-subject variation and assume random effects that gives some structure to error terms that characterises individual variation due to some factor levels. Finally, we will demonstrate non-linear statistical framework as a basis for estimation of population and individual viral dynamics parameters and how models may be used to draw biological relevant interpretations and aid clinical decision-making within the context of Limpopo HIV/AIDS data.

In Chapter 4, we will address the following objectives: we will analyse HIV/AIDS data using linear mixed effects models and Cox extended semi-parametric models; we will use shared parameter joint models to establish the strength of association between the longitudinal evolution of viral load and hazard rate to death; comparing separate and joint models, and various association measures such as parametric joint and shared parameter models approach; we will compare the average evolutions between males and females; showing how marker-specific evolutions are related to each other; computing prediction for time to death for any randomly selected HIV positive patient by considering patient's viral load; and come up with good joint model(s) that will handle simultaneously both the repeated measurements as well as the survival outcomes in the presence of clustering in the South African HIV/AIDS dataset.

1.3 Introducing the Limpopo Province HIV/AIDS Dataset

The dataset used in this thesis was a secondary data obtained from Limpopo Department of Health for the period 2001 to 2016. For the purpose of this study,more focus was on the data collected and recorded between 2011 January and 2017 January because of its reliability, since the data between 2001 to 2010 December were haphazardly captured and recorded, and secondly, there were no National Department of Health guidelines by that time. The structure of the longitudinally collected HIV/AIDS dataset is presented in Table 1.5.

Table 1.4 shows that Limpopo Department of Health has five districts, namely, Mopani District Municipality; Vhembe District Municipality; Capricorn District Municipality; Waterberg District Municipality and Sekhukhune District Municipality and twenty five local municipalities (Figure 1.1). Limpopo Department of Health has a total of 543 health care facilities. These facilities comprise of Clinics, Community Health Centres, District Hospitals, Regional Hospitals, and Provincial Hospitals. The total number of HIV patients that were included in our study were 9215 across Limpopo Province, of which 2776 (30.1%) died and 6439 (69.9%) were right censored. HIV/AIDS patients as defined by WHO were as follows: clinical stage I: 36.4%; clinical stage II: 36,5%; clinical stage III: 20.2%; and clinical stage IV: 6.9%. Furthermore, 6386 (69.3%) were females patients and 2829 (30.7%) males patients.

Table 1.4: Districts, Municipalities and number of Facilities in Limpopo Province.

District	Municipality	Number of health care facilities
Capricorn	Aganang	12
	Blouberg	25
	Molemole	9
	Lepelle-Nkumpi	26
	Polokwane	39
	Total	111
Mopani	Ba-Phalaborwa	12
	Greater Giyani	30
	Greater Letaba	23
	Greater Tzaneen	38
	Maruleng	13
	Total	116
Greater Sekhukhune	Ephraim Mogale	21
	Elias Motsoaledi	22
	Fetakgomo	16
	GreaterTubatse	31
	Makhuduthamaga	27
	Total	117
Vhembe	Makhado	54
	Musina	4
	Mutale	17
	Thulamela	56
	Total	131
Waterberg	Bela-Bela	5
	Lephalale	11
	Modimolle	5
	Mogalakwena	34
	Mookgophong	4
	Thabazimbi	9
	Total	68

Table 1.5: Limpopo Province HIV/AIDS Longitudinal Dataset

Id	Dst	Mn	Sex	Age	Event	PreOI	Cs	CD4	VL	T	StartDate	EndDate
1	Capricorn	Molemole LM	F	43	1	1	3	4.97	15.98	31	31/04/2012	23/07/2014
2	Mopani	Greater Giyani LM	F	56	1	1	4	7.82	16.35	42	12/07/2011	17/12/2013
3	Vhembe	Makhado LM	M	24	0	0	1	23.61	55.90	7	06/01/2011	11/07/2015
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
9213	Sekhukhune	Blouberg LM	F	44	0	1	2	3.34	342.78	20	21/02/2013	28/06/2013
9214	Waterberg	Mogalakwena LM	M	25	0	1	1	18.34	417.98	1	16/08/2014	26/01/2016
9215	Capricorn	Polokwane LM		M	1	1	4	3.88	667.89	14	12/09/2012	23/01/2013

In Table 1.5 the 1st column contain the unique HIV patient's identifier. The 2nd column contains 5 municipality districts (Dst) of Limpopo Province, and the 3rd column contains 25 local municipalities (Mn) corresponding to 5 districts. The 4th column gives the sex of HIV patients. The 5th column is the age of HIV patient at the 1st visit to the clinic. The 6th column is the status of an event, taking the value one if the HIV patient died and zero if censored. The 7th column gives the Previous Opportunistic Infection (PreOI) status at the start of HAART, taking the value 1 if the patient was on TB treatment, and zero otherwise. The 8th gives the AIDS clinical stage (Cs) of HIV patient based on WHO classification system for HIV infected patient: (Stage 1= Asymptomatic infection; Stage 2= Symptomatic infection; Stage 3= Advent opportunistic infections; Stage 4=ill-defining infections). The 9th column gives the square root of CD4 measurement at the baseline. The 10th column gives the natural logarithmic viral load (VL) measurements. The 11th column is the time taken for an event to occur since patient's 1st visit to the clinic.

1.4 Modelling HIV/AIDS Longitudinal Dataset

1.4.1 Univariate Survival Models

Survival is the phrase used to describe the analysis of a data that correspond to the time from a well-defined time origin until the occurrence of some particular event or endpoint. If the event is the death of a patient, the resulting data would be literally survival times. However, the data of similar form can be obtained when the event is not necessarily fatal, for an example, the relief of a pain, or the occurrence of symptoms. The methodology can also be applied to data from other application areas, such as survival times of animals in an experimental study, the time taken by an individual to complete a task in psychological experiment, etc., [17]. For the purpose of this study, survival analysis has been phrased in terms of the survival times of HIV/AIDS patients from the entry of study until event (death) occur.

Special features of survival data (Figure 1.3):

i) Firstly, survival data are generally not symmetrically distributed, but yield positively skewed histogram. As a consequence it would not be reasonable to assume that data comes from normal distribution. The challenge is circumvented by transforming the data to give a more symmetric distribution, for example by taking the square root, logarithms, hyperbolic sine, etc.;

ii) Secondly, survival data times of an individual is said to be censored if the endpoint of interest has not be observed at the end of study or individual has been lost to follow-up. Lost to follow-up can occur when individual after being recruited to clinical trial, a patient moves to another part of the country, and can no longer be traced. Patient can also be censored if death is from another cause (e.g., car accident) that is known to be unrelated to

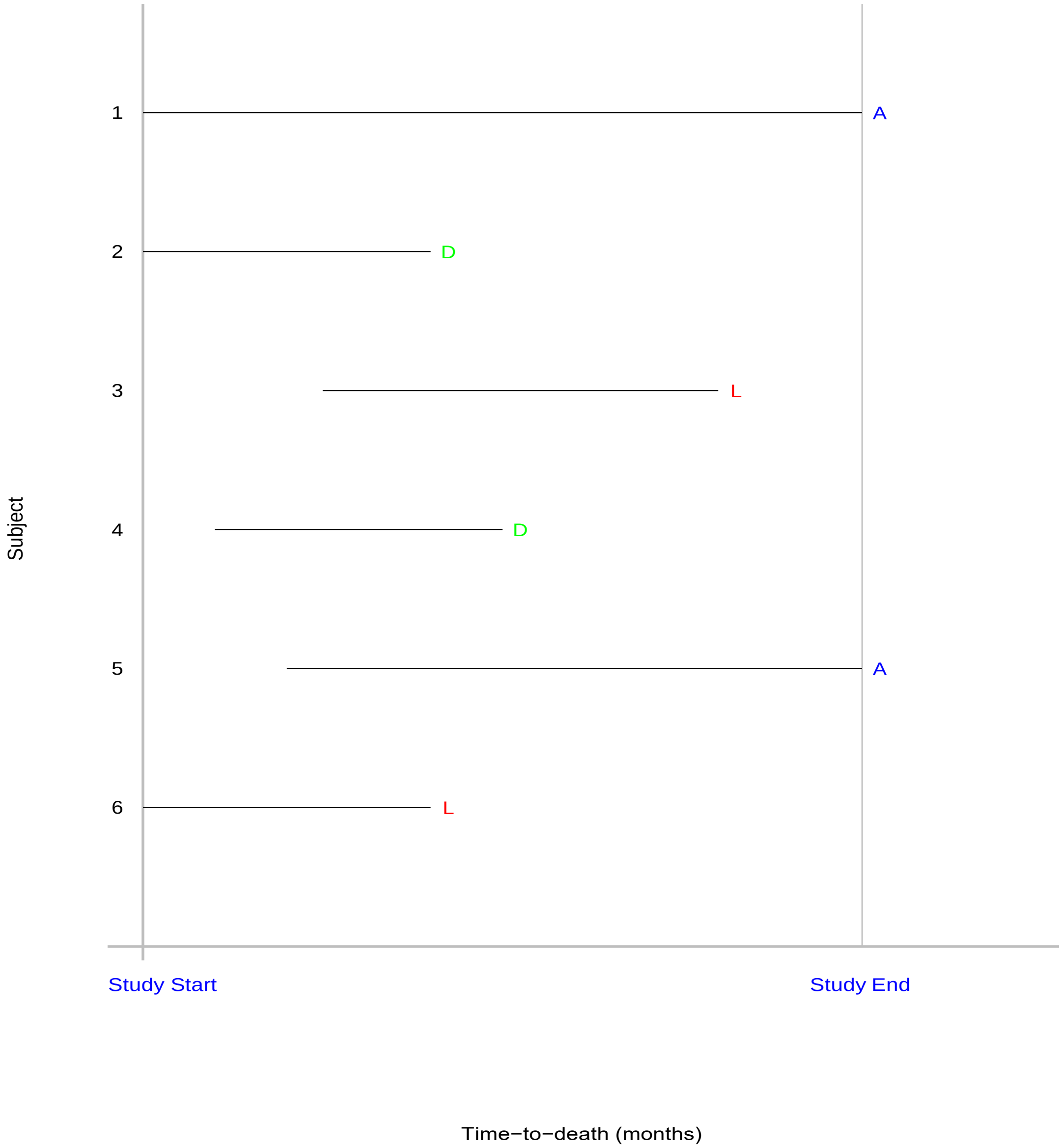


Figure 1.3: A sample of six HIV/AIDS patients with their time-to-death (in months). *L* refers to lost to follow-up from the study; *D* refers to the event death and *A* indicates that the person is still alive at the end of the study.

the treatment.

1.4.1.1 Censoring

In reliability as well as survival analysis, greater interested was placed on a non-negative random variable T ($T \geq 0$). This variable T can be discrete with values $\{0, 1, ...\}$ or continuous on $(0, \infty)$. Sometimes we cannot fully observe this random variable T but only observe some boundaries for this time. There is another variable $C(C \geq 0)$ which we call the censoring variable and which obscure the observation of T [18].

Survival censoring is broadly categorised into three types, namely:

i) Type I or fixed censoring;

ii) Type II censoring and

iii) Type III censoring or random censoring.

Type I or fixed censoring

Let $t_c \in R$ be a fixed time point and take a sample lifetimes $T_1, T_2, ..., T_n$. Hence we get a sample $Y_1, Y_2, ..., Y_n$, where

$$Y_i \;=\; \begin{cases} T_i, & \text{if } \; T_i \leq T_c, \\ t_c, & \text{if } \; T_i > T_c, \quad i = 1, 2, ..., n. \end{cases} \tag{1.1}$$

Example: The study stopped at a fixed time.

Type II censoring

Let $r < n$ with $r \in N$ and denoted by $T_1, T_2, ..., T_n$ the ordered lifetimes. We observe until the r^{th} system fails. Hence, we get

$$Y_i \;=\; \begin{cases} T_i, & \text{if } \; T_i \leq T_r, \\ T_c, & \text{if } \; T_i > T_r, \quad i = 1, 2, ..., n. \end{cases} \tag{1.2}$$

Example: Industrial test trial.

Type III censoring or random censoring

Let $C_1, C_2, ..., C_n$ be a sample of censoring times. We observed a sample couples $(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)$ where, for i=1,2,...,n.

$$Y_i = min(T_i, C_i) = \begin{cases} T_i, & \text{if } T_i \le T_{C_i}, \\ C_i, & \text{if } T_i > T_{C_i}, \quad i = 1, 2, ..., n. \end{cases} \tag{1.3}$$

$$\delta_i = I(T_i, C_i) = \begin{cases} 1, & \text{if } T_i \le T_{C_i}, \\ 0, & \text{if } T_i > T_{C_i}, \quad i = 1, 2, ..., n. \end{cases} \tag{1.4}$$

In general we assume that, for i=1,2,...,n; T_i and C_i are independent. Right censoring occurs when a subject leaves the study before an event occurs, or the study before the event has occurred. For example, when we consider HIV/AIDS patients in Limpopo Province, South Africa, and the study ends within five year (2011 January - 2016 January). Those patients who survived by the end of study (January 2016) were censored or those patients who left the study at time t_c and the event occurred (t_c, ∞) were also censored.

Left-censoring

We observe a sample $(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)$ where, for i=1, 2, . . . , n.

$$Y_i = max(T_i, C_i) = \begin{cases} T_i, & \text{if } T_i \ge C_i \\ C_i, & \text{if } T_i < C_i \end{cases} \tag{1.5}$$

$$\delta_i = I(T_i \geq C_i) = \begin{cases} 1, & \text{if } T_i \geq C_i \\ 0, & \text{if } T_i < C_i \end{cases} \tag{1.6}$$

Left-censoring occur when the actual survival time of individual is less than that observed. Left-censoring normally occur far less than right-censoring, and thus for the purpose of this study will be focusing on right-censored survival data. For example, the subject is said to be left censored if it is known that the failure occurs some time before the recorded follow-up period. For example, you conduct a study investigating factors influencing days to first oestrus in daily cattle. Population observation started at 40 days after calving but find that several cows in the group have already had an oestrus event. These cows are said to be left censored at 40 days.

Survival and hazard functions

Suppose that the random variable T has a probability distribution with underlying density function f(t). The distribution function is given by

$$F(t) = P(T < t) = \int_0^t f(u)du$$

$$F(t) = P(T < t) = \int_0 f(u)du$$

The survivor function, $S(t)$, is defined as follows:

$$S(t) = P(T \geq t) = 1 - F(t). \tag{1.7}$$

Hence, survival function, can be used to represent the probability that an individual survives from the time of origin to some time beyond t.

The hazard function $h(t)$ is defined as follows:

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{P(t \leqq T \leqq t+\delta t | T \geqq t)}{\delta t} && (1.8) \\ &= \lim_{\delta t \to 0} \frac{P(t \leqq T \leqq T \geqq t+\delta t)}{\delta t} [\frac{1}{P\,(T \geq t)}] \\ &= \lim_{\delta t \to 0} \frac{F(t+\delta t) - F(t)}{\delta t} [\frac{1}{S(t)}] \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

It follows from equation (1.8) that

$$h(t) = \frac{-d(logS(t))}{S(t)} \qquad (1.9)$$

and so

$$S(t) = exp[-H(t)] \qquad (1.10)$$

and

$$\begin{aligned} H(t) &= {}_{0}^{t} \lambda(u)du && (1.11) \\ H(t) &= {}_{0} \lambda(u)du && (1.11) \\ &= -logS(t) && (1.12) \end{aligned}$$

$H(t)$ is called cumulative hazard function.

Survival Likelihood

Due to the existence of censoring in survival data, the survival likelihood significantly different from classical likelihood for independent data without censoring. First, we will consider the construction of likelihood function for the right censoring. Next, we will give the general expression for the likelihood function accommodating left, right and interval censoring.

Suppose we have a random sample of size n from a specific population with independent survival times $T_1, T_2, ..., T_n$. However, due to censoring, we don't always have the opportunity observing these survival times. We denote C as the censoring process and $C_1, C_2, ..., C_n$ the censoring times. Thus the observed data are the minimum of the survival time and censoring time for each subject in the sample and the indication whether or not the subject is censored. We have the observed data (Y_i, δ_i), $i = 1, 2, ..., n.$, where $Y_i = min(T_i, C_i)$ is the time recorded, and δ_i indicates whether we observed an event or the subject was censored [17].

Let f (.)(F(.)) and g(.)(G(.)) denote probability density functions(distribution functions) for T an C, respectively. We assume that that T and C are independent.

For the right censored data with random censoring, the likelihood contribution of an event time (y_i, δ_i) is given by $(1 - G(y_i))f(y_i)$. On the other hand, for a right-censored observation $(y_i = c_i, \delta_i = 0)$ the contribution to the likelihood is given by $(1 - F(y_i))g(y_i)$. Hence, the likelihood is given by (owing to independence):

$$L = \prod_{i=1}^{n} [(1 - F(y_i))f(y_i)]^{\delta_i}[(1 - F(y_i))g(y_i)]^{1-\delta_i}$$

If we further assume that the distribution of the censoring times does not depend on the parameters of interest related to the survival function, called non-informative censoring [191, 20], the factor $(1 - F(y_i))^{\delta_i}$ and $(g(y_i))^{1-\delta_i}$ are not informative for inference on the survival function and, therefore, they can be deleted from the likelihood resulting in

$$\begin{aligned} L &= \prod_{i=1}^{n} (f(y_i))^{\delta_i} (S(y_i))^{1-\delta_i} \\ &= \prod_{i=1}^{n} (h(y_i))^{\delta_i} S(y_i) \qquad (1.13) \end{aligned}$$

or alternatively, with D the set of survival times and R the set of right censored times

$$L = \prod_{d \in D} f(y_d) \prod_{r \in R} (S(y_r))$$

The likelihood discussed above can be generalised to other types of censoring such as right, left and interval censoring. The likelihood expression for such data is given by

$$L = \prod_{i \in D} f(y_i) \prod_{i \in R} S(y_i) \prod_{i \in L} (1 - S(y_i)) \prod_{i \in I} (S(li) - S(r_i)) \qquad (1.14)$$

where D is the set of survival times, R is the set of right censored times, L is the set of left censored times and , I , is the set of interval censored times with l_i the lower limit and r_i the upper limit of the interval.

1.4.1.2 Modelling the hazard function

The basic model for survival data to be considered in this study is the proportional hazard model. This model was first proposed by Cox (1972) and also come to be known as the Cox regression model. The model is based on the assumption that the hazard of death at any given time for individual in one group is proportional to the hazard at any given time for a similar individual in the other group. This is the assumption of proportional hazards, which underlies a number of methods for analysing survival data. The model is therefore referred to as semi-parametric model since it consists of non-parametric baseline hazard function unspecified. The standard Cox proportional hazards model assumes a hazard function of the form:

$$h_i(t) = h_0(t)\ \exp(X_i(t)\beta),\ i = 1, \ldots, n. \qquad (1.15)$$

In equation (1.15) we assume the presence of covariates, where,

- t represents the survival time.
- $h(t)$ is the hazard function determined by a set of p covariates $(X_1, X_2, \ldots, X_p)$.
- the coefficients $(\beta_1, \beta_2, \ldots, \beta_p)$ measure the impact (i.e., the effect size) of covariates.
- the term $h_0(t)$ is called the baseline hazard function. It corresponds to the value of the hazard if all the X_i are equal to zero (the quantity exp(0) equals 1). The 't' in $h(t)$ reminds us that the hazard may vary over time.

There are several important assumptions for appropriate use of the univariate Cox proportional hazards models, including:

- independence of survival times between individuals in the sample.
- multiplicative relationship between the predictors and the hazard.
- constant hazard ratio over time.

Parametric Survival Models

When the Cox regression model is used in the analysis of survival data, there is no need to assume a particular form of probability distribution for survival times. On the other hand, if the assumption of a particular probability distribution for the data is valid, inferences based on such an assumption will be more precise. In particular, estimates of quantities such as relative hazards and median survival times will tend to have smaller standard errors than they would in the absence of a distributional assumption [17].

A probability distribution which plays a central role in the analysis of survival data is the Weibull distribution, introduced by W. Weibull in 1951 in the context of industrial reliability testing. The hazard function of the Weibull distribution is given by:

$$h(t) = \lambda t^{\gamma(\gamma-1)} \tag{1.16}$$

with $\gamma, \lambda > 0$, where λ is a scale parameter, and γ is a shape parameter.
In parametric proportional hazards models we assume a particular parametric function for the baseline function $\lambda_0(t)$. A particular assumption in the case of Weibull corresponds to:

$$h_0(t) = \lambda \gamma t^{\gamma-1}$$

When $\lambda = 1$ then the survival times has exponential distribution.

Accelerated Failure Time Models

The general form of an accelerated failure time (AFT) models is given by:

$$log(T) = \beta X + log(\tau) \tag{1.17}$$

where log(T) is the natural log of the time to failure event, βX is a linear combination of explanatory variables and log(τ) is an error term. Using this approach τ is the distribution of survival times when $\beta X = 0$. If we assume that τ follows a log-normal distribution, then log of survival times will have a normal distribution, which is equivalent to fitting a linear model to the natural log of survival time. Equation (1.17) can also be written as follows: $\tau = \exp(-\beta X)T$ or $ln(\tau) = -\beta X + log(T)$. Clearly, the linear combination of predictors in the model βX can act additively or multiplicatively on the log of time: they speed up or slow down time by a multiplicative factor. In this case $\exp(-\beta X)$ is called accelerated parameter such that if $\exp(-\beta X) > 1$ time passes more quickly, if $\exp(-\beta X) = 1$ time passes at a normal rate, and if $\exp(-\beta X) < 1$ time passes more slowly. There are other AFT, namely: log-normal, log-logistic, and gamma, which can be expressed as accelerated time models. The accelerated failure time coefficients represents the expected change in $\log(T)$ for one unit change in the predictor [25].

1.4.2 Mixed-Effects Models for Longitudinal Dataset

Linear mixed-effects models are important models that can be used to analyse correlated data. Such data include clustered observations, repeated measurements, longitudinal measurements, multivariate observations, etc. Linear mixed-effects models, like many other types of statistical models, describe a relationship between a response variable and some of the covariates that have been measured or observed along with the response. In mixed-effects models at least one of the covariates is a categorical covariate representing experimental or observational unit in the dataset. In medical sciences the observational units are often the human or animal subjects in the study. In Agricultural Sciences, the experimental units may be the plots of land or the specific plants being studied [31].

In all of these cases the categorical covariate or covariates are observed at the set of discrete levels. We may use numbers, such as identifiers, to designate the particular levels that we observed but these numbers are simply labels. The important characteristic of a categorical covariate is that, at each observed value of the response, the covariate takes on the value of one of a set of distinct levels.

If there is a set of possible levels of the covariate that is fixed and reproducible, l the covariate are modelled using fixed-effects parameters. If the levels that we observed represent a random sample from the set of all possible levels random effects in the model is incorporated [28].

The linear mixed-effects models (LME) are defined as follows:

$$\begin{aligned} \mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \varepsilon_i,\ i = 1, 2, .., n. \\ \mathbf{b}_i \sim N(\mathbf{0}, D), \\ \varepsilon \sim N(\mathbf{0}, \Sigma_i), \end{aligned} \tag{1.18}$$

where β_i=$(\beta_1, \beta_2, ..., \beta_p)^t$ is a $p \times 1$ vector of fixed effects, $\mathbf{b_i} = (b_{i1}, b_{i2}, ..., b_{iq})^t$ is a $q \times 1$ vector of random effects, the $n_i \times p$ matrix X_i and the $n_i \times q$ matrix Z_i are known design matrices may contain covariates, ε_i=$(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{in_i})^t$ represents random errors of the repeated measurements within-individual i (cluster), D is a $q \times q$ variance-covariance matrix of the random effects, and Σ_i is a $n_i \times n_i$ covariance matrix of the within-individual errors. We assume that $\Sigma_i = \sigma^2 I_{n_i}$ (homoscedastic conditional independence model) where I_{n_i} is the $n_i \times n_i$ identity matrix, i.e., the within-individual measurements are assumed to be independent with constant variance. The value of σ^2 represents the magnitude of the individual variations, and the value of Σ represents the magnitude of the between-individual variation, Wang and Taylor (2000) showed that LME model (1.18) is always identifiable if $\Sigma_i = \sigma^2 I_{n_i}$. The fixed effects β are population-level parameters and are the same for all individuals, while random effects b_i are individual-level, representing individual variation from population-level parameters. Since individual shares the same random effects, the multiple measurements within each individual or cluster are correlated. The LME model allows unbalanced data in the response which is an advantage of mixed models.

1.4.3 Joint Modelling of Longitudinal and Survival Dataset

To analyse the HIV/AIDS dataset in this study we will utilise the framework of joint models for longitudinal and survival data. The main idea behind these models is to couple a survival models for the continuous time-to-event process with mixed-effects model for longitudinal outcomes. The basic joint model is written as:

$$\begin{aligned} y_i(t) &= x_i^t(t)\beta + z_i^t(t)b_i + \varepsilon_i(t) \qquad (1.19) \\ h_i(t) &= h_0(t)\exp[\gamma^t w_i + a\{x_i^t(t)\beta + z_i^t(t)b_i\}], t > 0 \end{aligned}$$

where β denotes the vector with regression coefficients of the design matrix for the fixed effects $x_i^t(t)$ and $z_i^t(t)$ denotes the row vectors of the design matrix for the random effects b_i. In particular, the fixed effects describe the average longitudinal evolution in time, random effect describe how each patient deviates from this averaging evolution, where a quantifies the effect of the underlying longitudinal outcome to the risk for an event. If the value of $a = 0$, it means that there is no association between the longitudinal marker and the event time, which means that information from the longitudinal marker does not improve on the estimate of the survival time association effect a compared with an analysis based on the time to event alone. In that case no joint model is needed and one can ignore the longitudinal data when carrying out the survival analysis [72]. Moreover, it is assumed that the risk for the an outcome dependent event is associated with true an unobserved value of the longitudinal outcome (Figure 1.4).

Figure 1.5 of basic form joint models assumes that the hazard function at any particular time point t, denoted by the vertical dashed line, is associated with the value of the longitudinal process (green line) at the same time point. The blue line represents the assumption behind the time-dependent Cox model,

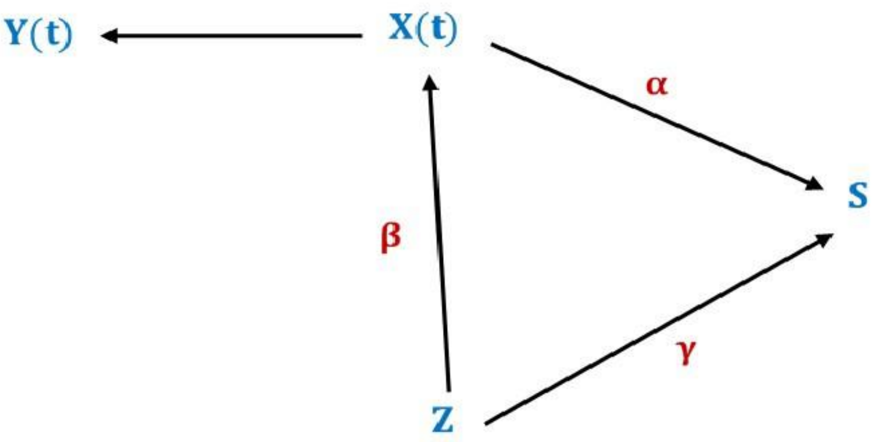


Figure 1.4: A diagram showing the relationships between Y(t), observed longitudinal data; X(t), trajectory function; S, survival; Z, treatment; γ, treatment effect on survival; β, treatment effect on longitudinal process; α, effect of longitudinal process on survival.

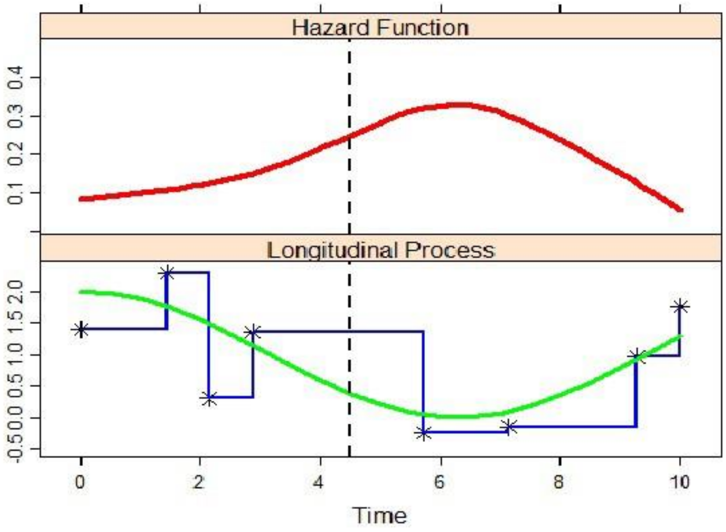


Figure 1.5: Joint Models for Longitudinal and Survival Outcomes.
Source: Rizopoulos et. al, (2012)

which posits that the value of longitudinal outcome remain constant between observation times. The framework of joint models can be used to account for both endogenous time-varying covariates and non-random dropout. Estimation of the joint model is based on the joint distribution of the two outcomes, and can be done either under maximum likelihood or under a Bayesian approach [172].

[redacted]

Chapter 2

Survival Analysis of HIV/AIDS Dataset

2.1 Introduction

In many clinical trials randomization is often used to evaluate new treatments for patients with human immune virus (HIV), and, in these HIV trials clinical progression of AIDS is used as the primary outcome. In our studies we will use viral load measurements as a marker for HIV. The recent development of assay techniques for quantifying HIV-1 RNA in HIV-infected patients makes it possible to use viral load (HIV-1 RNA copies) as surrogate marker to accelerate AIDS clinical trials. Prentice (1989) defined a good surrogate marker that it should have the following three properties:

- The marker should be related to prognosis.
- The distribution of values for the marker should be different for individuals receiving an effective treatment versus those receiving a placebo.
- The beneficial effects of a good treatment should be mediated through its effect on the marker. That is, patients with the same value of a marker should have the same prognosis whether they are receiving prognosis or a

placebo. Clearly in such a case the better prognosis associated with a good treatment could be explained by the change in the value of the marker for that treatment.

Some deterministic HIV-1 dynamics models have been develop to describe the interaction between HIV and its host cells in individual patients [152, 154, 153, 151]. A stochastic model was also proposed [155]. The Bio-mathematicians and biologists developed models which were too complicated because they contained too many unknown parameters to be used in the analysis of real clinical data [156]. Recently, more simplified models have been proposed and applied to real virological data from clinical trials [157, 152]. These studies of HIV dynamics have led to a new and a far better understanding of the pathogenesis of HIV infection. These HIV dynamics models provide a global better picture of virus elimination and production process during antiretroviral treatment (ART) for each individuals treatment. Hence, in evaluating the efficacy of anti-HIV treatments and understanding HIV infection pathogenesis it is of critical importance to estimate viral dynamics parameters for the whole population and as well as each individual patient in Limpopo Province, South Africa.

2.2 HIV/AIDS Dataset in Limpopo Province

The data of HIV/AIDS patients were collected at 543 health care facilities across Limpopo province by the Limpopo Department of Health. These facilities comprise of district hospitals, regional hospitals, provincial hospital, community health centres, specialised sites, non-medical sites and clinics. The database contains the following variables for each patient: patient's identifier, viral load, CD4 cell counts measurements, prescribed ART, gender, previous opportunistic infection, AIDS clinical stages, district, different types of health care facilities, etc. In this secondary data, there were 9215 patients diagnosed with HIV between January 2011 and January 2017. The majority of these patients were females accounting for 69.3% of the total HIV/AIDS patients in Limpopo Province. Mopani district accounted for 36.53% of 9215 HIV/AIDS patients. At the end of this period (30.1%) HIV/AIDS patients died and (69.9%) were rightly censored. These censored HIV/AIDS patients were either lost to follow-up, or migrated to other provinces or died at the end of the study.

2.3 Aim of the study

The aim of the study is to assess the risk factors associated with mortality rate among HIV positive patients in Limpopo Province, and how these risk factors contributes to high death rate in Limpopo Province.

2.3.1 Objectives

We will be addressing the following objectives in this chapter:

i) to use Kaplan-Meier to compare the average evolutions between gender, districts, health care facilities, previous opportunistic infections, and AIDS clinical stages;

ii) to analysis survival data using both Cox proportional hazard and parametric models; and

iii) to compare the semi-parametric and parametric models.

2.4 Methods of Survival Data Analysis

2.4.1 Non-parametric estimation of survival function

Survival analysis is a collection of statistical procedures for the analysis of data in which the outcome variable of interest is time until an event occurs. By event we mean death, disease incidence, relapse from remission, or any designated experience of interest that may happen to an individual. When doing survival analysis, we usually refer to time variable as survival time. The capital T is denoted as a random variable for a person's survival time, and small letter t as any specific value of interest for random variable T.

In most survival analyses a key data analytical problem, called censoring is considered. In essence, censoring occurs when we have some information about individual survival time, but we do not know the time exactly. Most survival time data are right-censored, because the true survival time interval, which we do not really know, has been cut off (i.e., censored) at the right side of the observed time interval, giving us true survival time.

We let δ be random variable defined as follows:

$$\delta = \begin{cases} 1, & \text{if } T \le C, \\ 0, & \text{if } T > C. \end{cases}$$

which means that an individual is either died or censored. An individual who does not die, that is, does not get an event during study period, must have been censored either before or at the end of the study.

The survival function $S(t)$ gives the probability that a random variable T exceeds the specific time t and defined as follows

$$S(t) = P(T > t)$$

This is the probability that an event will not occur by time t. Kaplan and Meier (1958) developed an estimator for the survival function $S(t)$, given by

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i})^{\delta_i}$$

where:

- d_i=number of patients died at t_i;
- n_i=number of patients at risk before t_i.

This $\hat{S}(t)$ is called the Kaplan-Meier or product-limit estimator. The product-limit estimator is based on an assumption of non-informative censoring. Non-informative censoring means that knowledge of censoring time for an individual provides no further information about the person's likelihood of survival at a future time had the individual continue with the study. The Kaplan-Meier (KM) estimator provides an efficient means of estimating the survival function for right-censored data. Some properties for the KM estimator can be found in [161].

In this study we will apply Greenwood's estimator(1926) of the variance of Kaplan-Meier estimator, which is defined by:

$$Var(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{2.1}$$

A pointwise $100(1 - \alpha)\%$ confidence interval for the survival function $S(t)$ at time t_0 (t_0 is fixed time), termed the linear confidence interval is defined by $\hat{S}(t_0) \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}ar(\hat{S}(t_0))}$.

2.4.2 Log-rank test

Log-rank test is a hypothetical test to compare the survival distributions of two or more samples. It is a non-parametric test and appropriate to use when data are right skewed and censored and censoring must be non-informative.Therefore, the whole survival function under the null hypothesis of different groups,would be compared.

$$H_0 : S_1(t) = ... = S_K(t), 0 < t < \tau$$

Since the true survival functions are unknown in each group, the non-parametric test would be opted for.
Firstly, let us consider a simpler test where $K = 2$, then

$$H_0 : S_1(t) = S_2(t), \quad 0 < t < \tau$$

Furthermore, suppose we observed, from the populations $j = 0, 1$ $(T_{j1}, \delta_{j1}), (T_{j2}, \delta_{j2}),$ $..., (T_{jn_j}, \delta_{jn_j})$.
Under H_0, both populations are equal.

Let the lifetimes $\tau_1, \tau_2, ..., \tau_k$ be k-ordered, distinct death times. Now, at the $£-th$ death time, a 2×2 contingency table is as follows:

Population	Yes (Death)	No (Alive)	Total
0	d_0	$n_0 - d_0$	n_0
1	d_1	$n_1 - d_1$	n_1
Total	d	$n_l - d$	n

where d_j is the number of deaths and n_j is the number at risk in population j at this time. Under null hypothesis H_0 and conditional on the marginals, d_j has a hypergeometric distribution. That is, $d_j \sim$ Hypergeometric(n , d , n_j).

Therefore, the mean and the variance of d_j is given by

$$E(d_j) = \frac{n_1\, d}{n}$$

with

$$Var(d_j) = \frac{n_1\, n_0\, d_j(n - d)}{n^2(n-1)}$$

Suppose that the sample size at each death time is very large, by the central limit theorem the hypergeometric distribution can be approximated by a normal distribution.

Let us assume that the contingency tables at different death times are independent, then the log-rank test is given by

$$T = \frac{[\ {}^{k}_{=1}(d_1 - \frac{n_{1\pounds}\, d_{\pounds}}{n_{\pounds}})]^2}{{}^{k}_{=1}\ \frac{n_{1\pounds} n_{0\pounds} d_{\pounds}(n_{\pounds} - d_{\pounds})}{n_{\pounds}^2 - 1}}$$

which is, under H_0 is approximately x^2 distributed with $df = 1$. The log-rank test is a special case of the Terone-Ware class of tests, where

$$T = \frac{[\ {}^{k}_{=1}\, w\, (d_1 - \frac{n_{1\pounds}\, d_{\pounds}}{n_{\pounds}})]^2}{{}^{k}_{=1}\ w^2\, \frac{n_{1\pounds}\, n_{0\pounds}\, d_{\pounds}\, (n_{\pounds} - d_{\pounds})}{n^2_{\pounds} - 1}}$$

and where $w \geq 0$ are weights.

Test	w
Log-rank	1
Wilcoxon	n
Peto-Peto	$\tilde{S}(t_i)$
Harrington-Fleming (p,q)	$\tilde{S}(t_i)^p(1 - \tilde{S}(t_i))^q, p, q \geq 0$

Let $\tilde{S}(t) = \cap_{t_i \leq t}(1 - \frac{d_i}{n_i+1})^{\delta_i}$ In a practical data analysis, the choice of weights is extremely important for the following reasons [28]:

- Log-rank test has optimal power to detect alternatives in which hazards are proportional;
- Wilcoxon test is more ”sensitive” to ”early” differences in survival curves; and
- Harrington-Fleming test with $p = 0$, $q > 0$ is ”sensitive ” to ”later” differences. Hence, it is recommended in literature that the choice of appropriate test should be made before seeing the data.

To get desired trend survival curves for three or more ordered groups are compared. Thus study compared four AIDS clinical stages of HIV/AIDS patients using log-rank test for the trend:

$H_o : S_1(t) = S_2(t) = \ldots = S_K(t), \quad t \leq \tau.$
$H_A : S_1(t) \geq S_2(t) \geq \ldots \geq S_K(t), \ t \leq \tau$ at leat one >.

Then the test statistic is [17]:

$$T = \frac{\sum_{j=1}^{K} a_j(O_j - E_j)}{\sqrt{\sum_{j=1}^{K} \sum_{k=1}^{K} a_j a_k V_{jk}}}$$

$$\sim \mathrm{N}\ (0, 1)\ .$$

under H_0, where $a_1 < a_2 < \ldots < a_K$ are ordered of scores and

$$O_j = \sum_{=1}^{k} d_j$$
$$E_j = \sum_{=1}^{k} \frac{n_j\, d}{n}.$$

The test for trend has higher statistical power than the normal log-rank test [17].

2.4.3 Semi-Parametric Cox Proportional Hazard Models

In order to study the potential impact of covariates on time to event of that particular process, a number of modelling approach has been developed, including proportional hazard (PH) model developed by Cox (1972). This proportional hazard model proved exceedingly useful in the analysis of survival data. This model specifies that the hazard function for the survival time T associated with $p \times 1$ vector of covariates $\mathbf{X}$ takes the form:

$$h(t|\mathbf{X}) = \mathbf{h_0(t)exp(X}^t\beta) \tag{2.2}$$

where $\beta_1, ..., \beta_p$ are regression coefficients, $h_0(t)$ (baseline hazard) is the underlying hazard function, which is unknown and unspecified non-negative function of time, that does not depend on thecovariates.

This model has two kinds of assumptions that require verification before one can rely on statistical inferences and predictions the model yields. The first assumption is that the relationship between log hazard (log cumulative hazard) and a covariate is linear. The strength of this model is that $h_0(t)$ is left unspecified (unknown function). It represents the hazard of an individual with covariates equal to zero. The second assumption is the time independence of the covariates in the hazard function, that is, the ratio of the hazard function for two individuals with different regression vectors does not vary with time (the PH assumption).

To estimate the coefficient $\beta_1, \beta_2, ..., \beta_p$, Kalbfleisch and Prentice (2011) proposed a partial likelihood function based on a conditional probability of failure, assuming that there are no tied values in the survival times. Although the partial likelihood is not a full likelihood, the estimators obtained from this maximisation have been shown to be consistent and asymptotically normal.

However, in practice, tied survival times are common. Throughout the study the Breslow approximate likelihood function which assumes piecewise hazard between failure times [28, 29, 30] is applied to estimators for $\beta_1, \beta_2, \ldots, \beta_p$.

Cox-Snell residuals

The Cox-Snell residual procedure is the most widely used in the analysis of survival data, and was first proposed by Cox and Snell (1968). The Cox-Snell results for the *ith* individual, $i = 1, \ldots, n$ is given by

$$r_{cs_i} = \exp(\hat{\beta}_i^t x_i) \hat{H}_0(t_i) \qquad (2.3)$$

where $\hat{H}_0(t_i)$ is the estimated cumulative hazard function at time t_i. If the Cox's regression model is satisfied, we get that r_i is a censored sample of an exponential distribution with lambda equal to one.

2.4.4 Parametric Survival Models

Cox's semi-parametric model is the most frequently employed regression for survival data, however, parametric models may offer some advantages [28, 57]. Based on asymptotic results, Efron (1997) and Oakes (1997) showed that, under certain circumstances, parametric may lead to more efficient parameter estimates than the Cox model. When empirical information is sufficient, parametric model models can provide some insight into the shape of the baseline hazard. Secondly, extrapolation of survival functions become possible, which, although speculatively, may be of interest to the applications. Fully parametric models involve stronger assumptions than semi-parametric models. The challenge is to choose the appropriate parametric model because there is danger of mis-specification of the model, hence, the statisticians tend to prefer Cox model.

There are a number of parametric models that are frequently employed in survival analysis to describe and model event times. In parametric proportional hazards model, a particular parametric distribution for the baseline hazard $h_0(t)$ of equation 2.2 is assumed. The distributions which are commonly used for survival time are Weibull, exponential, log-logistic, log-normal and generalised gamma. An assumption for the baseline hazard corresponds to:

$$h_0(t) = \lambda\gamma(\lambda t)^{\gamma-1} \tag{2.4}$$

with $\gamma > 0$, $\lambda > 0$, and γ is a shape parameter which allows the density to take a variety of shapes, depending on the value of the shape parameter, λ is a scale parameter and it provides information on the way the hazard is stretched out.

When making this parametric assumption in equation 2.4 for the baseline hazard, it follows that the event times are Weibull distributed. The shape parameter works in the following ways:

- if $\gamma < 1$, then the hazard is monotonically decreasing with time.
- if $\gamma > 1$, then the hazard is monotonically increasing with time.
- if $\gamma = 1$, then the hazard is flat and have the exponential model. That is, Weibull nests the exponential model.

The survival function for the Weibull is:

$$S(t) = e^{-(\lambda t)^{\gamma}} \tag{2.5}$$

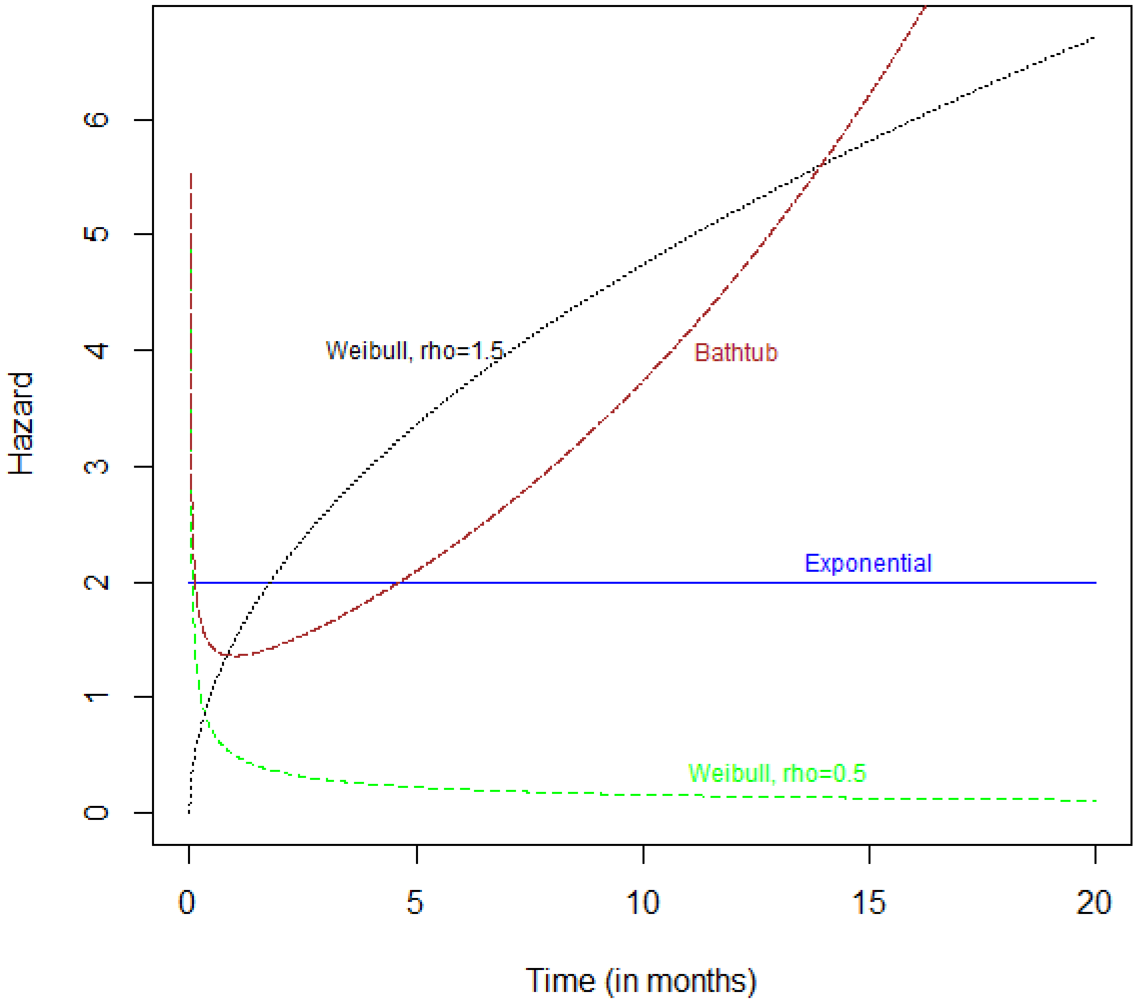


Figure 2.1: Different types of hazard functions that are often encountered in practice. Here our $\lambda = 2$.

and the density function is:

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1} e^{-(\lambda t)^{\gamma}} \tag{2.6}$$

using , Weibull likelihood is given by:

$$L = \prod_{i=1}^{N} [\lambda\gamma(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}}]^{d_i}[e^{-(\lambda t)^{\gamma}}]^{1-d_i} \tag{2.7}$$

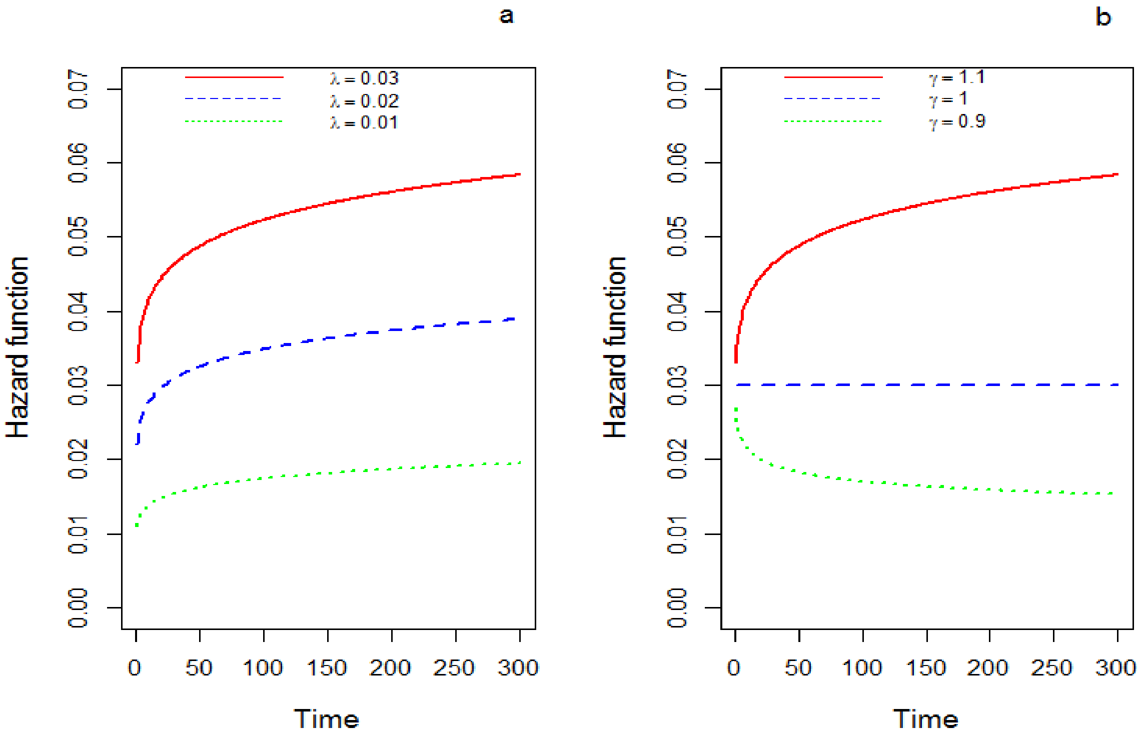


Figure 2.2: Weibull hazard functions with (a) different scale(fixed $\gamma = 1.1$) and (b) different shape parameters (fixed $\lambda = 0.03$).

An important aspect of the Weibull distribution is therefore its proportional hazards property: Weibull distributed event times with the same parameter γ lead to the proportional hazard model. Weibull distributed event times are often used in practice because they seem to be able to describe the actual evolution of the hazard function in an appropriate way in manycircumstances.

2.4.4.1 Accelerated Failure Time Models (AFT)

The accelerated failure time model is an alternative if the proportional hazards assumptions does not hold. Different diagnostic tests have been developed to evaluate the proportional hazard assumption [35]. In contrast to the proportional hazard model, the accelerated failure time model is the best characterised in terms of the survival function [32]. In the presence of covariates, the accelerated failure time model can be written as follows :

$$h_i(t) = \exp(x_i^t\beta)\, h_0(t\exp(x_i^t\beta)) \tag{2.8}$$

When the parametric proportional model is Weibull, the baseline hazard function is given by

$$h_0(t) = \lambda\gamma(\lambda t)^{\gamma-1} \tag{2.9}$$

and the event times follows Weibull distribution. As a results, the hazard given covariates for this parametric model can be written as:

$$h_i(t) = \lambda\gamma t^{\gamma-1}(\exp(x_i^t\beta))^{\gamma} \tag{2.10}$$

and the survival function is given by:

$$S_i(t) = \exp(-\lambda\gamma^{-1}\exp(\gamma x_i^t\beta)). \tag{2.11}$$

and the density function is

$$f_i(t) = [\exp(-\lambda t^{\gamma}\exp(\gamma x_i^t\beta))][\lambda\gamma t^{\gamma-1}\exp(\gamma x_i^t\beta)] \tag{2.12}$$

Thus, all subjects have Weibull distributed event times with the same shape parameter but different scale parameters (Figure 2.2).

2.4.4.2 Log-linear models

Instead of modelling the hazard functions, we can model the survival time directly. The log-linear model is an example of such modelling, and it is given by:

$$logT_i = \mu + \mathbf{x}_i^t a + \sigma E_i \qquad (2.13)$$

where T_i the event time for subject i, μ the intercept, $\mathbf{x}_i$ the vector of covariates for subject i, a the vector containing the covariate effects, σ the scale parameter, and E_i the random error term for subject i. The error term is assumed to have a fully specified distribution.

In model (2.13), let us assume that E_i (error term) has Gumbel distribution, that is,

$$E_i \sim \exp(e - \exp(e)) \quad -\infty < e < \infty \qquad (2.14)$$

and $\exp(E_i) \sim \exp(1)$ (exponential law with mean one).
The model in (2.13) can be written in terms of survival function as follows [32]:

$$\begin{aligned} S_i(t) &= P(T_i > t_i) = P(logT_i > logt_i) \\ &= P(\mu + x_i^t a + \sigma E_i > logt_i) \\ &= P(E_i > (logt - \mu - \mathbf{x}_i^t a/\sigma) \\ &= P[\exp(E_i) > \exp((logt - \mu - \mathbf{x}_i^t a/\sigma] \\ &= \exp[-\exp((logt - \mu - \mathbf{x}_i^t a)/\sigma)]. \end{aligned}$$

The last expression can be re-written as

$$S_i(t) = \exp[-\exp(\frac{-\mu}{\sigma})t^{\frac{1}{\sigma}}\exp(x_i^t(\frac{a}{\sigma}))]. \qquad (2.15)$$

The Weibull accelerated failure time can be written in terms of the survival function

$$S_i(t) = \exp(-\lambda t^{\gamma} \exp(\rho \mathbf{x}_i^t \beta). \qquad (2.16)$$

when we compare equation (2.15) and (2.16), clearly the two models correspond as follows:

$\lambda = \exp(\frac{-\mu}{\sigma})$ $\qquad \gamma = \sigma^{-1}$ $\qquad \beta = -\alpha.$

On the other hand, the survival function for the Weibull proportional hazards model is given by

$$S_i(t) = \exp(-\lambda t^{\gamma} \exp(\mathbf{x}_i^t \beta)). \qquad (2.17)$$

Again, comparing (2.15) and (2.17), clearly the two models corresponding with

$\lambda = \exp(\frac{-\mu}{\sigma}),$ $\qquad \gamma = \sigma^{-1},$ $\qquad \beta = \frac{\alpha}{\sigma}.$

thus

$$\beta = (-\alpha) * (\sigma).$$

Therefore, the parameter estimates from the log-linear model can be easily be transformed into parameter estimates for either the Weibull accelerated failure time model or Weibull proportional hazards model. For the proportional hazard model Duchateau (2007) derived the variance of a ratio $(-\alpha) * (\sigma)$ of two parameter estimates as follows:

$$Var(\hat{\beta}_i) = Var(\frac{\hat{\alpha}}{\hat{\sigma}}).$$

First, Duchateau (2007) used delta-method to approximate the variance as follows:

Let us consider

$$\boldsymbol{\zeta}^t = (\zeta_1, \ldots, \zeta_k)$$

and a univariate continuous function $g(\boldsymbol{\zeta})$. Now the Taylor expansion $g(\boldsymbol{\zeta})$ is given by

$$g(\boldsymbol{\zeta}) \approx g(\boldsymbol{\zeta}) + \gamma^t(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \qquad (2.18)$$

where $\hat{\zeta}$ is the maximum likelihood estimator of ζ with

$$\gamma^t = (\frac{\partial(\boldsymbol{\zeta})}{\partial\zeta_1}, \ldots, \frac{\partial(\boldsymbol{\zeta})}{\partial\zeta_k})$$

being the vector of the first partial derivatives evaluated at $\boldsymbol{\zeta}$. From 2.18 we obtained

$$Var[g(\hat{\zeta})] \approx \boldsymbol{\gamma}^t Var(\hat{\zeta}), \boldsymbol{\gamma}$$

where $Var(\hat{\zeta})$ is the variance-covariance matrix of $\hat{\boldsymbol{\zeta}}$. For the specific case of the Weibull proportional hazards model we have

$$\boldsymbol{\zeta}^t = (\mu, \boldsymbol{\alpha}, \sigma)$$

and

$$\boldsymbol{\beta}_j = \frac{-\alpha_j}{\sigma}$$

We therefore have

$$\zeta^t = (0, 0, \ldots, \frac{-1}{o}, \ldots, 0, \frac{\alpha_j}{\sigma^2}).$$

Given many zeros, it is easy to observe that

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_j) &= \boldsymbol{\gamma}^t Var(\hat{\zeta})\boldsymbol{\gamma} \\ &= \frac{1}{\sigma^2} Var(\hat{\alpha}_j) - 2\frac{\alpha_j}{\sigma^3} Cov(\hat{\alpha}_j, \hat{\sigma}) + \frac{\sigma^2}{\sigma^4} Var(\hat{\sigma}) \end{aligned} \quad (2.19)$$

An estimate for $Var(\hat{\beta}_j)$ is obtained by using in (2.19) as estimates for $Var(\hat{\alpha}_j)$, and $Cov(\hat{\alpha}_j, \hat{\sigma})$ the corresponding entries of the inverse of the observed information matrix and by replacing α_j and σ in by their corresponding estimates $\hat{\alpha}_j$ and $\hat{\sigma}$ obtained by fitting the log-linear models.

Other Parametric survival Models

In addition to Weibull distribution, other flexible and popular distributions for time to event data are log-logistic, log-normal and generalised gamma.

2.4.4.3 Log-logistic

A variable T has a log-logistic distribution if its logarithm follows the logistic distribution in (2.13) with density function

$$f_i(t) = \exp \frac{\exp(a)\kappa t^{\kappa-1}}{(1+\exp(a)t^{\kappa})^2} \tag{2.20}$$

and survival function

$$S_i(t) = \frac{1}{1+\exp(a)t^{\kappa}} \tag{2.21}$$

and hazard function

$$\lambda_i(t) = \frac{\kappa t^{\kappa-1}\lambda^{k}}{1+(t\lambda)^{\kappa}} \tag{2.22}$$

with $a \in \mathrm{R}$ and $\kappa > 0$.

2.4.4.4 Log-normal

Again, a variable T has a log-normal distribution if its logarithm follows the normal distribution in (2.13) with density function given by

$$f_i(t) = \frac{1}{t\sqrt{(2\pi\gamma)}} \exp[-\frac{1}{2\gamma}(\log(t)-\mu)^2],\ tER \tag{2.23}$$

and survival function

$$S_i(t) = 1 - \Phi(\frac{\log(t)-\mu}{\sqrt{\gamma}}),\ tER \tag{2.24}$$

with $\mu \in \mathrm{R}$ and $\gamma > 0$ and hazard function $\frac{f(t)}{S(t)}$.

2.4.4.5 Generalised Gamma

The generalised gamma models has a quite complicated specification involving two shape parameters . The density of the generalised gamma distribution is:

$$f(t) = \frac{\lambda\gamma(\lambda t)^{\gamma\kappa-1}\exp[-(\lambda t)^{\gamma}]}{\Gamma(\kappa)} \tag{2.25}$$

where $\lambda_i = \exp(-(X_i))$ is a scale parameter and ρ and κ are two shape parameters. The two parameters allow for quite a flexible hazard rate, including a U- shape. The elegance characteristic of the generalised gamma model is that it nests several of the other parametric models as special cases: Weibull, exponential, log-normal, and the standard gamma. Thus, these models are good for adjudicating between competing models. The shape parameter works as follows:

- if $\kappa = 1$, the Weibull distribution is implied.
- if $\kappa = \gamma = 1$, the exponential is implied.
- if $\kappa = 0$, the log-normal is implied.
- if $\gamma > 0$, the gamma distribution is implied.

2.4.4.6 Maximum likelihood estimation for θ

In survival analysis, some observations are censored, hence, estimation has to be adapted to censoring. Suppose we have a non-informative censored sample $(Y_1, ..., Y_n)$ where

$$Y_i = min(T_i, C_i) \quad \delta_i = I(T_i \leq C_i), i = 1, ..., n.$$

with

- a sample $T_1, ..., T_n \sim f(t, \theta)$ lifetimes, where f is known and θ unknown. We denote the survival function by $S(t, \theta)$.

- a sample $C_1, ..., C_n \sim g(c)$ censoring times with survival function $G(t)$.

- T_i and C_i are independent.

For an uncensored observation ($\delta = 1$), the contribution to the likelihood is given by

$$\begin{aligned} P(Y \leq y,\ \delta = 1) &= P(min(T,\ C) \leq y,\ T \leq C) \qquad (2.26) \\ &= P(T \leq y, C \geq T) \\ &= {}_{o}^{y}\, G(t)f(t)dt \\ &= {}_{o}\, G(t)f(t)dt \end{aligned}$$

$$\Rightarrow f_{Y,\delta=1}(y, \theta) = f(y, \theta)G(y)$$

For a censored observation ($\delta = 0$), similarly we get

$$f_{Y,\delta=0}(y,\ \theta) = g(y)S(y,\ \theta)$$

Hence we get the likelihood function

$$L(\theta) = \prod_{i,\delta=1} f(y_i, \theta)G(y_i) \times \prod_{i,\delta=0} g(y_i)S(y_i, \theta)$$

Since we assumed that the censoring is non-informative, we get

$$\begin{aligned} L(\theta) &= \prod_{i,\delta=1} f(y_i, \theta) \times \prod_{i,\delta=0} S(y_i, \theta) \qquad (2.27) \\ &= \prod_{i=1}^{n} f(y_i, \theta)^{\delta_i} S(y_i, \theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} h(y_i, \theta)^{\delta_i} S(y_i, \theta)^{1-\delta_i} \end{aligned}$$

In practice, we always have covariates present, as a result we get similar expression for the likelihood:

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(t_i|x_i,\theta)^{\delta_i} S(t_{i|x\,i,\theta})^{1-\delta_i} \qquad (2.28)\\
&= \prod_{i=1}^{n} \lambda(t_i|x_i,\theta)^{\delta_i} S(t_i|x_i,\theta)
\end{aligned}
$$

We will get estimates for different parameters when we maximise this $L(\theta)$, and the maximum likelihood estimate for θ is asymptotic normal. In the full non-parametric case, we assume the Cox proportional hazard model $\lambda(t|\mathbf{x} = \lambda_0(t)\exp(\mathbf{x}^t\boldsymbol{\beta})$. We now wish to estimate the regression parameter $\boldsymbol{\beta}^t = (\beta_1, \ldots, \beta_q)^t$. It is well known that the Cox partial likelihood function is given by

$$
L(\beta) = \prod_{j=1}^{n} \left(\frac{\exp^{\mathbf{x}_i^t \boldsymbol{\beta}}}{\sum_{k\in n_j} \exp^{x_k^t \boldsymbol{\beta}}} \right)^{\delta_i} \qquad (2.29)
$$

To estimate β , $L(\beta)$ is maximised. Although the partial likelihood is not a full likelihood, the estimators obtained from this maximisation have been shown to be consistent and asymptotically normal.

2.5 Comparison of Cox PH & Parametric Models

The proportional models are routinely employed for analysis of time-to-event data in medical research in the presence of covariates [178], however, parametric models may offer advantages. If the assumption of proportional hazard (PH) is violated, then, the results from a PH model will be difficult to generalise to situations where the length of follow-up is different to that used in the analysis. Furthermore, it would be difficult to translate the results into the effect upon the expected median duration of illness for a patient in a clinical setting [54]. The second disadvantage of the PH model is that the underlying hazard function is common across all patients, for example, the hazard functions for any two patients with baseline h_0 vector $\mathbf{x}_1$ and $\mathbf{x}_2$ are constrained to be proportional and the method of estimation is based on this, whereas, the method of estimation depends critically on evolving risk sets through time [55]. The PH model is considered to be semi-parametric and as such has advantage of being able to cope with variety of basic shapes for the common hazard function across patients.

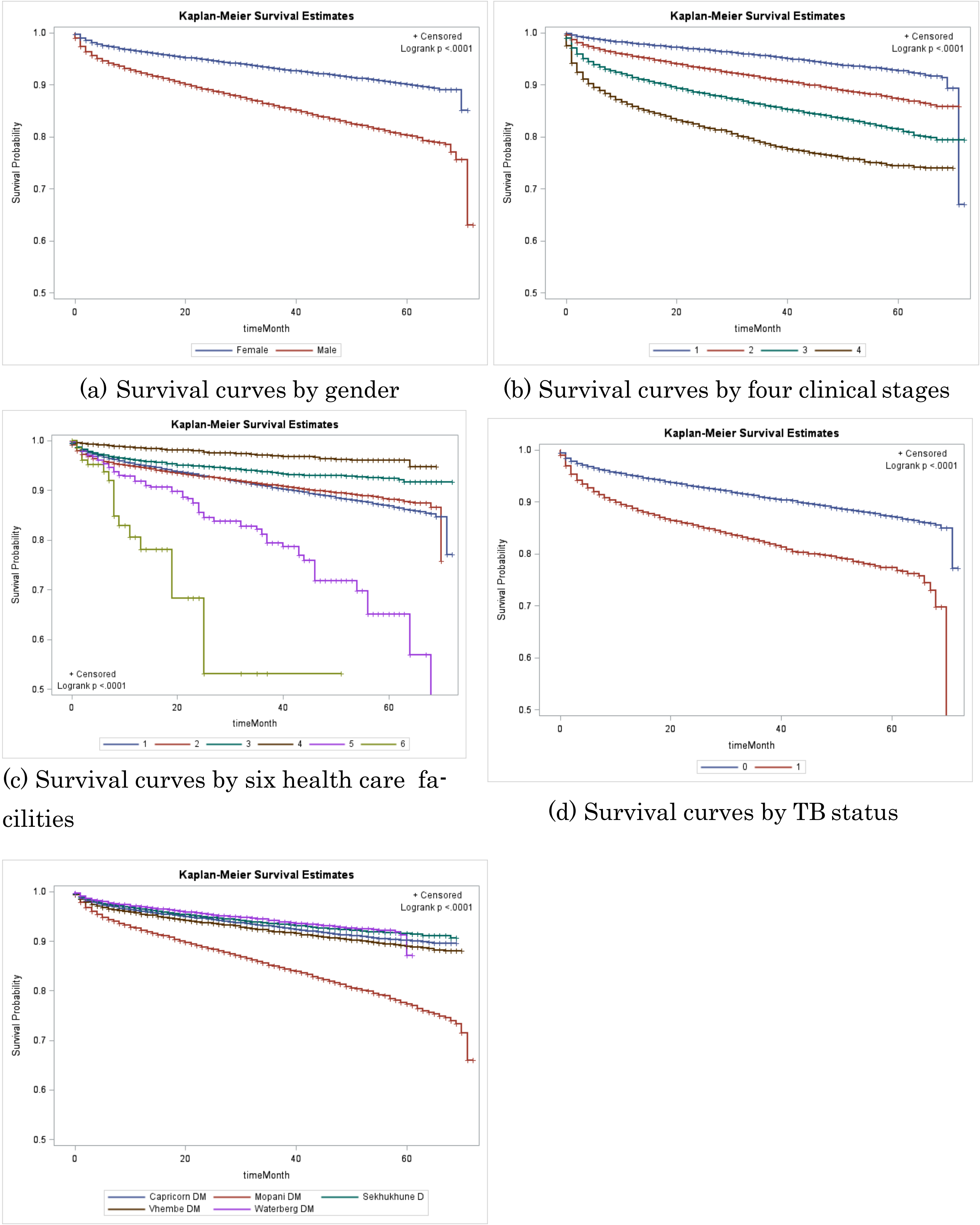
The accelerated failure time (AFT) approach is an alternative strategy for the analysis of time to event data and can be suitable even when hazards are not proportional and this family of models contains a certain form of PH as a special case. The results of AFT model may be easier to interpret and more relevant to clinicians, as they can be directly translated into expected reduction or prolongation of median time to event, unlike the hazard ratio. Based on asymptotic results, Efron (1997) and Oakes (1997) showed that, under certain circumstances, AFT approach leads to more efficient parameter estimates than PH models. With decreasing sample size, relative efficiencies may further change in favor of AFT models. When empirical information is sufficient, AFT models can provide some insight into the shape of the baseline hazard. Furthermore, extrapolation of survival functions becomes possible [56].

2.6 Results

2.6.1 Results from Cox Proportional Hazard Models

The survival curves for HIV/AIDS patients in Limpopo Province are presented in Figure 2.3. The plots displayed distinct separation between curves, reflecting strong negative effects on the main effects, that is, gender, AIDS clinical stages, different types of health care facilities, TB status of a patient at ART initiation, age at baseline and five districts on the survival probabilities.

In estimating these survival curves in univariate Cox proportional hazard model,



(a) Survival curves by gender

(b) Survival curves by four clinical stages

(c) Survival curves by six health care facilities

(d) Survival curves by TB status

(e) Survival curves by five districts.

Figure 2.3: KM Survival curves by different covariates.

one covariate at the time was fitted in the model (Cox PH) while other covariates were fixed at zero or held constant; for example, when a KM curve for gender covariate was plotted, covariates such as districts, Previous Opportunistic Infection (PrevOI), clinical stages, health care facilities, CD4 counts, age at baseline, were all fixed at zero, so that survival curves for males and females can be compared effectively [121, 23].

Sub-figure 2.3a shows that there is a significant difference in survival times between males and females [log-rank $P< 0.0001$], hence, the null hypothesis of no difference in survival times is rejected in favour of the alternative. The curves of females are above that of males, hence, it is concluded that females have significant longer survival times as whole. Therefore, morbidity and mortality were lower in females that in males.

Sub-figure 2.3b shows survival curves of HIV/AIDS patients in Limpopo Province for the four AIDS clinical stages of HIV as describe by World Health Organisation (WHO). What is noticed first in Sub-figure 2.3b is that the curves are distict and parallel. Secondly, there is a significant difference in survival times between these four AIDS clinical stages [log-rank $P< 0.0001$]. The survival curve of clinical stage I is far above all other survival curves, with clinical stage IV the lowest. Hence, it is concluded that patients in clinical stage I have significantly the longest survival times, whereas patients in clinical stage IV have the lowest survival times. This is in agreement with similar findings by [80, 81, 82]. Therefore, the patients' mortality rate were the highest in clinical stage IV and lowest in clinical stage I.

Sub-figure 2.3c provides information about the survival times differences of HIV/AIDS patients in various health care facilities that are available in Limpopo Province. The survival curves show there is a significant difference in survival times of patients in these six health care facilities [log-rank P< 0.0001]. it is noticed that survival curves of clinics and community health care centres & district hospitals are crossing (i.e. indicating non-proportional hazards) at the later time as a result of survival times having greater variance in patients' care than the others. Since the crossing happens at the later stage time, weighted log-rank test using Harrington-Fleming method which is sensitive to later differences will be preferred (see section 2.5.2). The survival curve of Provincial Hospital is far above all survival curves, with survival curve of of Specialised Site at the bottom, and the test was significant. It is therefore concluded that patients attending health care at Provincial hospital have the longest survival times than any other health care facilities, with Specialised Site having patients' with shortest survival times. Furthermore, it is evident that morbidity and mortality were lowest in Specialised Sites than any other health care facilities.

Sub-figure 2.3d shows that there is a significant difference in survival times between patients who experienced TB before they were diagnosed as HIV positive and patients who were TB free [log-rank P< 0.0001]. Thus, the null hypothesis of no difference is rejected in favour of the alternative. Furthermore, the survival curve of patients who experienced TB before diagnosed with HIV is below that of patients who were TB free, hence, we conclude that patients who were TB free during diagnosis survive longer than those who experienced TB before. According to Suchindran et al., the risk of death in TB-HIV co-infected individual is double as compared with HIV infected individuals without TB. Other studies also report that the presence of TB co-infection is associated with higher mortality among HIV patients taking ART [60, 104, 103]. Therefore, mortality was higher with patients who experienced TB before than TB free patients.

Sub-figure 2.3e shows the survival curves of HIV/AIDS patients for five districts in Limpopo Province. There is a significant difference in survival curves in these five districts [log-rank $P < 0.0001$]. The Mopani district curve lies well below all other district curves, with Waterberg well on top of the others. Hence, the null hypothesis of no difference in survival times is rejected in favour of the alternative. Thus, it is concluded that patients in Waterberg district have significantly longer survival times as a whole, with Mopani district patients shortest survival times. As a result, the morbidity and mortality in Mopani district is highest than any other district.

Table 2.1: Cox PH model for clinical stages, age, and interaction of clinical stages by age of HIV patients in Limpopo Province, utilising Breslow method in handling ties.

Covariate	Parameter	Estimates	Standard Error	Wald Chi-Square	P-value
Clinical stage II	β_1	0.7606	0.07208	111.4010	< 0.0001
Clinical Stage III	β_2	1.70076	0.06592	665.6235	< 0.0001
Clinical Stage IV	β_3	2.1109	0.07785	735.4055	< 0.0001
Age	β_4	0.02575	0.00125	397.9842	< 0.0001
Clinical Stage II × age	β_5	-0.0091	0.00172	1.2376	< 0.2659
Clinical Stage III × age	β_6	-0.01159	0.00160	52.4507	< 0.0001
Clinical Stage IV × age	β_7	-0.00995	0.00187	28.3638	< 0.0001

Table 2.2 suggests that there is an interaction between patients in AIDS clinical stages III-IV and age. That is, the survival times for HIV+ patients depend on age in AIDS clinical stages III and IV because a local test for $\beta_6 = \beta_7 = 0$ will be rejected (P-value < 0.0001). Therefore, mortality rates for HIV+ patients in AIDS clinical stage III and IV dependent on age as compared with AIDS clinical stage I. Furthermore, Table 2.2 also suggests the mortality rate in AIDS clinical stage II patients by looking at a local test $\beta_5 = 0$ (P-value =0.2659) will not be rejected. That is, the effect of AIDS clinical stage II on survival times for HIV-infected patients is the same for different age groups as compared with clinical stage I. Hence the interaction of clinical stage II by age may be dropped

from the full model because the relative risk does not depend on age.

Table 2.2 gives the details of the Cox proportional hazard model fit. The important covariates in predicting survival time were age (years), CD4 counts, gender, clinical stages of HIV, districts, previous opportunistic infections, and different types of health care facilities. The hazard for a patient one-year older is 1.02 times that of a patient younger, suggesting that an increase in age shortens survival time, and it is highly significant ($P < 0.0001$). The relative risk of CD4 for HIV patients is 0.88 times that of lower CD4 count, showing that one-cell/μL increases for a patient's results in a longer survival times, and it is highly significant. The hazard ratio for males of HIV patients is 2.169 times than that of female patients, suggesting that the mortality rate of male is approximately twice that of female throughout the observed period, and it is highly significant ($P < 0.0001$). The estimated hazard ratios of death for Capricorn, Mopani, Sekhukhune and Vhembe districts are 1.236, 2.595, 1.092, and 1.411, respectively, as compared to that of Waterberg district, suggesting that Capricorn , Mopani and Vhembe districts have 24%, 160% and 41% higher hazards of death, respectively, than that of Waterberg district, and are all highly significant ($P < 0.0001$). The Sekhukhune district has 9% higher hazard of death as compared to that of Waterberg district, and it is moderately significant ($P<0.030$).

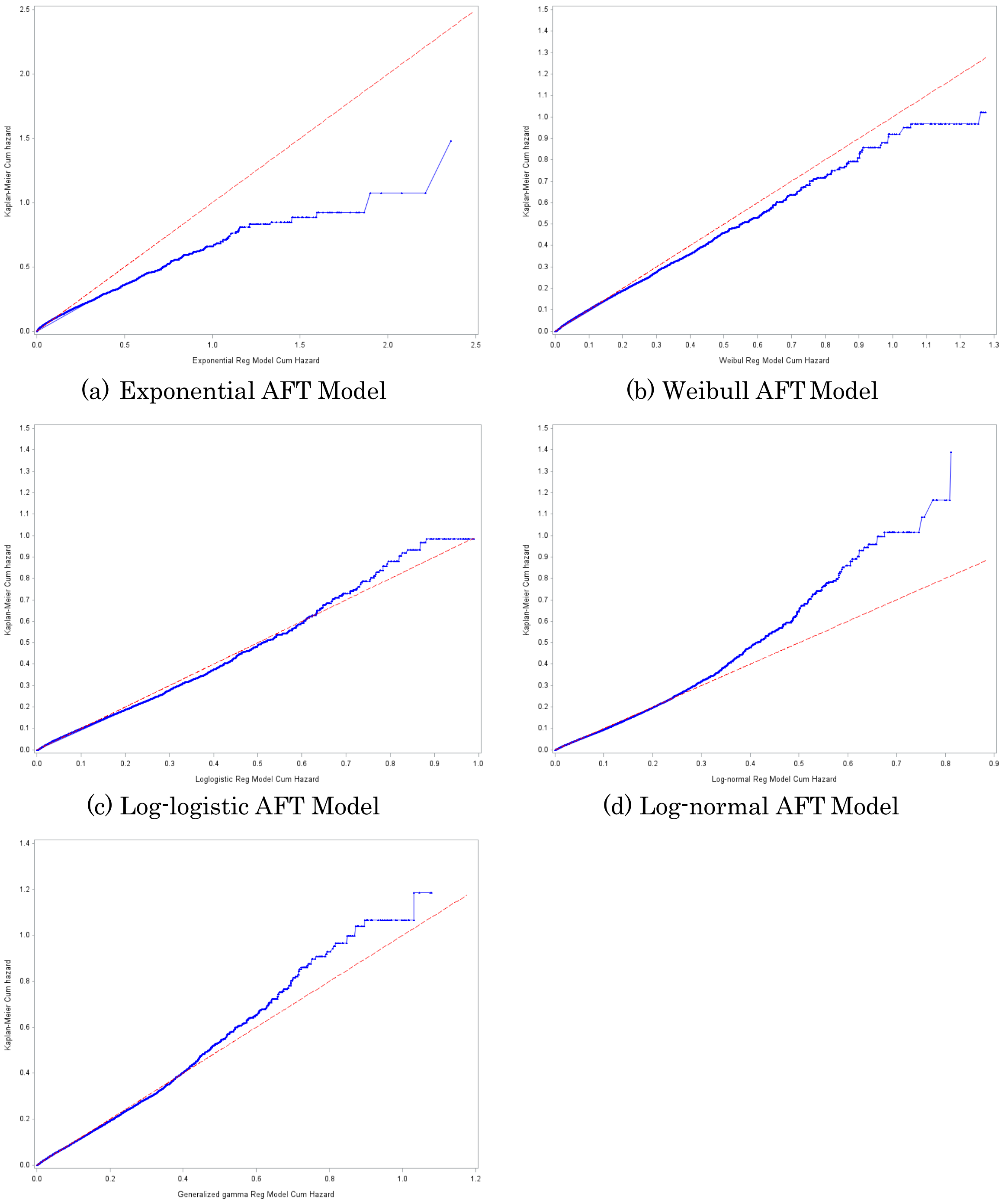
The estimated hazard ratios of death for clinical stage II, III, and IV are 2.142, 3.710 and 5.996, respectively, as compared to that of clinical stage I. Showing that the clinical stage II, III, and IV have roughly 114%, 271% and 500% higher hazard of death, respectively, as compared with clinical stage I, and are all highly significant ($P < 0.0001$). The estimated hazard ratio of death for patients who experienced tuberculosis (TB) before diagnosed with HIV-infection is 2.12, suggesting that patients who experienced TB before have roughly 112%

higher hazard of death as compared to those who were TB free, and it is highly significant. The hazard ratios of death for Regional and Provincial Hospitals are 0.7356 and 0.3072 respectively as compared to that of Clinics & Mobile Clinics, suggesting lower hazard of death by 26.4% and 69.3% for Regional and Provincial Hospitals respectively, as compared to that of clinics, and are all highly significant ($P < 0.0001$). The hazard ratios of death for Community Care Centre, Non-Medical Sites, and Specialised Sites are 1.0479, 1.8116, and 3.6921 respectively, as compared to clinics. That is an evidence that Community Care Centre, Non-Medical Sites, and Specialised Sites have higher hazard of death by approximately 5% , 81.2% and 269.2% respectively, as compared to that of clinics, and are all highly significant.

2.6.2 Model diagnostics

Figure 2.4 shows the Cox-Snell residuals plots from fitting the exponential, Weibull, log-logistic , log-normal, and generalised gamma models respectively with covariates gender, age at baseline, CD4 counts at baseline, districts, type of health care facilities, clinical stages, and previous opportunistic infections. We fitted these parametric AFT models in Figure 2.4 using Proc Lifereg SAS version 9.2 procedure. The Exponential and Log-normal models are not fitting well as they deviate from the straight lines, that is, sub-figure 3.1 and sub-figure 2.4d, respectively. The three other graphs (Weibull, log-logistic, and generalised gamma models) look similar and all are close to a straight line with unit slope and zero intercept. There is no significant difference observed with these three graphs. The results are similar to those obtained in table 2.5 (Akaike Information Criterion of five distributions fitted to the full model).

The difference among these three distributions are small with log-logistic distribution being slightly better than the others, however, in Table 2.5 generalised gamma model was a preferred model because of the its lowest AIC value



(a) Exponential AFT Model

(b) Weibull AFT Model

(c) Log-logistic AFT Model

(d) Log-normal AFT Model

(e) Gamma distribution

Figure 2.4: Cox-Snell residual plot for different parametric AFT model

and that is consistent with the findings of [59] . The small AIC value could have been influenced by one or more parameter as compared with log-logistic distribution. A paper by [50] suggested that Weibull is the best fitted parametric model for predicting survival following a diagnosis of AIDS and could be used for future projections of death from HIV/AIDS patients in Limpopo province for right censored data. However, based on our findings log-logistic model will be our final preferred model for our right censored data.

2.6.3 Results from Parametric Models

Many parametric models are accelerated failure time (AFT) models rather than proportional hazard models. The standard parametric models are exponential, log-logistic, log-normal and Weibull. The underlying assumption in these AFT models is that the effects of covariates act multiplicative with respect to survival time [28]. For Weibull distribution the AFT assumption holds, hence in Table 2.4, a one-unit difference in the age of a patient corresponds to a hazard ratio of $\exp(-.00445/\gamma) = \exp(-0.007295) = 0.99273$. The Weibull model suggests that the hazard of death for any given patient in Limpopo Province is approximately 0.99 times that for a patient one year younger, and it is highly significant ($P < 0.0001$). The patient CD4-counts hazard ratio is $\exp(0.000295/\gamma) = \exp(0.0004836) = 1.00048$, that suggests that the hazardof death of any given patient is roughly 1.00 times for a patient with a CD4 count one-cells/μL higher, and it is highly significant($P < 0.0001$). The hazard ratio for a male patient is 0.6495, shows that hazard of death for any male patient is 0.6495 times that for female patient, and it is highly significant ($P < 0.0001$).

In Table 2.3 the hazard ratios of patient in Mopani, Sekhukhune, Vhembe and Waterberg districts are 0.5359, 0.9561, 0.9165, and 1.0974, respectively, that suggest that the hazard of deaths of patients in these three districts are 0.54 times, 0.96 times, 0.92 times higher than that of Capricorn district and are all highly significant ($P < 0.0001$). While Waterberg is 1.1 times lower than that of

Capricorn district, and it is highly significant ($P< 0.0001$). The hazard ratios of patients in Clinical stage II, III and IV are 0.20025, 0.06161, and 0.02373, respectively, which suggest that the death hazard of patients for clinical stage II,III, and IV is 0.20 times, 0.06 times and 0.02 times higher than that of clinical stage I, and it is highly significant ($P< 0.0001$). The hazard ratio of patients who had experienced TB treatment in the past is 0.71800, hence, the hazard for patients who experienced TB treatment in the past is 0.72 times lower than those who never experienced TB treatment before HIV was diagnosed, and it is highly significant ($P< 0.0001$).

In Table 2.3 the hazard ratios of patients in Community Health Centres, District Hospitals, Regional Hospitals, Provincial Hospital and Non-Medical Sites are 1.5135, 0.9595, 1.2484, 1.7970, and 0.6863, respectively. There is a sufficient evidence that suggest that hazard of death of patients in Community Health Centres, Regional Hospitals and Provincial Hospital are 1.51times, 1.25times and 1.80 times, respectively, lower than that of clinics, and are highly significant ($P< 0.0001$). While the hazard ratios of patients in District Hospitals, Non-Medical Sites are 0.96times and 0.69times, respectively lower than that of clinics, and are highly significant ($P< 0.0001$).

There is a sufficient evidence that suggests that hazard of death of patients in Community Health Centres and Provincial hospital are 1.5 times and 1.80 times, respectively, higher than that of the clinics, and are highly significant ($P< 0.05$). There is also a sufficient evidence that suggests that hazard of death of patients District Hospitals and Non-Medical Sites are 0.96 times and 0.69times, respectively, lower than that of the clinics, and are statistically non-significant ($P> 0.05$). Finally, there is a sufficient evidence that suggests that hazard of death of patients the Regional Hospitals are 1.2 times higher than that of clinics, however, they are statistically not significant ($P> 0.05$). Since the Weibull distribution is skewed, a more appropriate and more tractable summary of the location of the distribution is the median survival time. In our study, the population median survival time is $[\ln 2/\lambda]^{\frac{1}{\gamma}} = 14.4$ months, while the population mean survival time is $\lambda^{\frac{-1}{\gamma}}\,\Gamma(1 + \frac{1}{\gamma}) = 102.3$ months.

Table 2.2: Hazard ratios from Cox proportional hazard model for HIV/AIDS dataset in Limpopo Province.

Univariate Analysis					
Covariate	Parameter Estimate(β_i)	Std. Error (β_i)	HR [$exp(\beta_i)$]	95% CI (HR)	P-value
Age	0.0073463	0.0013362	1.0073733	(1.0047;1.0100)	< 0.0001
CD4	-0.0004894	0.0001171	0.9995107	(0.9993;0.9997)	< 0.0001
GENDERMale	0.4290153	0.0396670	1.5357445	(1.4209;1.6599)	< 0.0001
Female (Ref)	0.0000		1.0000		
Mopani DM	0.6165256	0.0506981	1.8524805	(1.6773;2.0460)	< 0.0001
Sekhukhune DM	0.0448817	0.0605008	1.0459041	(0.9290 ; 1.1776)	0.030
Vhembe DM	0.0887012	0.0664949	1.0927541	(0.9592;1.2449)	< 0.0001
Waterberg DM	-0.0849743	0.0783749	0.9185359	(0.7877; 1.0710)	
Capricorn DM (Ref)	0		1.0000		
Clinical Stage II	0.76163	0.02240	2.142	(2.050;2.2383)	< 0.0001
Clinical Stage III	1.31109	0.02082	3.710	(3.562;3.865)	< 0.0001
Clinical Stage IV	1.79108	0.02416	5.996	(5.719;6.287)	< 0.0001
Clinical Stage I (Ref)	0.0000		1.0000		
PrevOI	0.3270792	0.0614804	1.3869114	(1.2295;1.5645)	< 0.0001
No PrevOI (Ref)	0.0000		1.0000		
Community Health centres	-0.4124507	0.0783964	0.6620258	(0.5677;0.7720)	0.0046
District Hospitals	0.0639858	0.0612065	1.0660773	(0.9456;1.2020)	0.0046
Regional Hospitals	-0.2353223	0.1309825	0.7903161	(0.6114;1.0216)	< 0.0001
Provincial Tertiary Hospital	-0.5871451	0.2006980	0.5559121	(0.3751;0.8238)	< 0.0001
Non-Medical Sites	0.3807131	0.1921661	1.4633277	(1.0041;2.1326)	< 0.0001
Clinics (Ref)	0.0000		1.0000		

[1] HR: Hazard ratio.
CI: Confidence Interval.

Comparison of Cox Proportional Hazard and parametric models Results

Table 2.4 compared the Cox proportional hazard and five parametric AFT full models of HIV/AIDS patients in Limpopo Province. The Akaike Information Criterion (AIC) reveals that parametric AFT models generally it fit the data better. The AIC and Cox-Snell residual graph (Sub-figure 2.4e) shows that generalised gamma model fitted our Limpopo Province HIV/AIDS data the best. Hence, Parametric regression models demonstrate better performance as compared with the Cox model for identifying risk factors for prognosis with Limpopo Province data. Our results is in agreement with Adelian *et al.*, (2015) and Teshnizi and Ayatollahi, (2017) while research done by Saikia and Barman, (2016) found that Cox PH model was better than other parametric counterparts for esophagus cancer patients data.

Table 2.3: Hazard ratios from Weibull proportional hazard model for HIV/AIDs patient's dataset in Limpopo Province.

Univariate Analysis				
Covariate	Parameter Estimate(β_i)	Std. Error (β_i)	HR	P-value
intercept	4.738667	0.05570		< 0.0001
Age	-0.004450	0.00082	0.99273	< 0.0001
CD4	0.000295	0.00007	1.00048	< 0.0001
Male	-0.263290	0.02440	0.64946	< 0.0001
Female (Ref)	0.0000	1.0000		
Mopani DM	-0.380558	0.03140	0.535867	< 0.0001
Sekhukhune DM	-0.027365	0.03690	0.95613	< 0.0001
Vhembe DM	-0.053213	0.04060	0.916463	< 0.0001
Waterberg DM	0.056707	0.04780	1.09742	< 0.0001
Capricorn DM (Ref)	0.00000	1.0000		
Clinical Stage II	-0.981000	0.01320	0.20025	< 0.0001
Clinical Stage III	-1.700001	0.01520	0.06161	< 0.0001
Clinical Stage IV	-2.28200	0.01730	0.02373	< 0.0001
Clinical Stage I (Ref)	0.0000	1.0000		
PrevOI	-0.202083	0.03750	0.71800	< 0.0001
No PrevOI (Ref)	0.0000	1.0000		
Community Health Centres	0.252808	0.04790	1.5135	< 0.0001
District Hospitals	-0.041380	0.03730	0.9595	< 0.0001
Regional Hospitals	0.135336	0.07980	1.2484	< 0.0001
Provincial Tertiary Hospital	0.357543	0.12300	1.7970	< 0.0001
Non-Medical Sites	-0.229605	0.11700	0.6863	< 0.0001
Clinics (Ref)	0.0000	1.0000		

[1] HR: Hazard ratio.
Ref : Reference.
Scale(ρ)=0.61

Table 2.4: Akaike Information Criterion of six distributions fitted to the full model.

Model	Log-likelihood	Number of covariates	Number of parameters	AIC
Exponential	-43318.18	16	1	86670.36
Weibull	-42573.93	16	2	85183.86
Log-normal	-42654.61	16	2	85345.21
Log-logistic	-42508.35	16	2	85052.70
Generalised gamma	-42422.84	16	3	84883.69
Cox's PH	-44340	17	2	107234.13

[1] AIC: Akaike Information Criterion

The HIV/AIDS patient data in Limpopo Province

The adequacy of the five parametric models was assessed (each with all covariates included) and their Akaike Information Criterion (AIC) values in Table 2.4 was presented. The generalised gamma model has highest log-likelihood than the other models and the lowest AIC, indicating that this distribution may be the most accurate and the best fitted model.

2.7 Discussion

This study managed to determine the risk factors associated with HIV/AIDS patients in Limpopo Province, using parametric AFT models(viz. exponential, Weibull, log-normal, log-logistic and generalised gamma) and Cox proportional hazard model. The AFT models provided generally a better description of the data than Cox proportional model [70, 55]. We found that female patients had a better survival time than their male counterparts, which is consistent with Alioum *et al.*, (2010) and Taylor-Smith *et al.*, (2010) findings. However, Ramafedi et al., (1995) found that gender does not have any significant difference on survival time of HIV/AIDS patients, which is contrary to our findings. In our study a decrease in the CD4 cell counts resulted in shorter survival time for HIV/AIDS patients, which is in agreement with the findings of other the researchers in the literature. The CD4 counts were found to be an important prognostic marker of HIV/AIDS patients. The patients with CD4 counts less than 200 cells were more likely to die than patients with more than 350 cells per mm^3, [60, 61, 64, 65, 66]. The results of our study had a strong inclination for the generalised gamma AFT model, better than others based on AIC. Earlier papers by Nakhaee et al., (2011) suggested that Weibull is the best fitted AFT model for predicting survival after a patient has been diagnosed HIV infection. Our findings suggested that prognostic factors(viz. age(years), CD4 counts, WHO clinical stages, districts, previous opportunistic infection, type of facilities, and gender) were statistically significant ($P < 0.0001$). Most previous studies suggested that age is a significant prognostic factor and according to Bachani et al., (2010), Ghate at al., (2011) and Kee (2009) a younger person undergoing ART is more likely to survive longer as compared with an older person, that is, old age is associated with high risk of disease progression.

Sub-figure 2.3(a) shows that there is gender difference in survival among men and women patients. A similar study was conducted by Cornell *et al.*, (2012) in South African context, and they found that these gender differences were due to immunologic and virologic response to treatment. Secondly, men are more likely to be lost to follow-up than women. Thirdly, HIV-infected men have higher mortality on ART than women in South African programmes, but these differences are only explained by more advanced HIV disease at the time of ART initiation [75, 76, 74, 77].

Researchers are in agreement that dealing with human rights, especially women rights could be one of the solutions of reducing the prevalence of HIV/AIDS. Moreover, research concerns that women rights underlies most health care problems in developng countries. Therefore tackling human rights especially the improvement of women status would be another way of dealing with the fundamental reality. Empowering them has long been seen as an important public health goals [78, 79]. Where women are independent, society tend to be much healthier than would otherwise be expected, because it is usually women who fight for better service and living conditions for their families. New researchers in South Africa suggested that women empowerment is possible, and it can be done by designing programmes that offered small business loans to South African women living in impoverished rural villages in Limpopo, one of South African's poorest provinces. Similar micro-finance programmes have helped poor women in many developed countries from Bangladesh to Brazil to gain a degree of independence by setting up small enterprises such as buying and selling food, clothes or cosmetics. That programmes can be of value since it allows women to participate in economic activities [74]. These programmes in Limpopo Province may rekindle ambition and purpose in community long demoralised by lack of opportunities, discriminatory laws, and culture of inequality that had numbed many poor people into dependency upon government welfare.

Limpopo Province is the third poorest Province in South Africa, and is predominately rural. The unemployment rates are very high as in 2015 the unemployment was 20.1%. This study shows that Mopani district has the highest HIV/AIDS patients death rate (HR= 1.85248). Owing it to the fact that Mopani district is in a rural setting and there are no employment prospects, and as a result, active individuals in economic activities tend to go out in search for jobs. Studies done by Collinson *et al.*, (2006) purported that male migration leads to high risk of sexual behaviour. Mining industry like Phalaborwa Mining Company in Mopani district is an important sector for migrant labour in Mopani district, but other industries like construction and security have also become increasingly important employment sectors which are found in urban areas. As aresults of migration, infected men return home to infect their partners over weekends, Easters, and Christmas holidays.

Other studies done in Limpopo province by Posel (2005) showed that rural residents had moderate amount of knowledge related to HIV, many rural residents had misconceptions and myths about HIV. They believe that AIDS is a new form of other longstanding illness, which traditional healers can cure, and is called tindzhaka (an illness caused by the breach of cultural taboos on sex during mourning period) . Many studies also showed that people had some knowledge about HIV/AIDS, particularly young people, who usually have access to mass media, and they also receive HIV education at school are worse affected. Sub-figure 2.4e, Table 2.2 and Table 2.3 show that mortality rate in Mopani district is very high compared to other four districts. A report from Joint United Nations Programme on HIV/AIDS and [49] indicated that rural populations might not be knowledgeable about HIV as was reported, because of high illiteracy levels, low level of education and poor infrastructure in the rural areas. These factors impeded access to health information as well as resources needed to prevent HIV infection. Study done by Mabunda *et al.*, (2015) in Limpopo showed that, in spite of the AIDS awareness campaign going onin South Africa, some segments of the population do not get the message , specifically uneducated in rural areas. The researcher suggested support groups within the rural community as an effective method of educating these people about HIV/AIDS [118].

This study shows that death rate of patients in the fourth clinical stage was the highest as compared to other stages. In the fourth clinical stage it is where the patient's vital organs (eg., kidneys, heart) start collapsing, and patient experiences HIV wasting syndrome, pneumocystis pneumonia, Kaposi sarcoma and symptomatic HIV-associatednephropathy.

Limpopo province recorded in 2008 a TB mortality rate of 12.4% with majority of deaths recorded among the economically active age group (24-54 years). The mortality was significantly associated with older age, extra pulmonary site of disease, HIV co-infection, smear negative pulmonary tuberculosis (PTB) and previous history of TB [118]. The association between HIV infection and TB was evident in this present study: HIV positive patients were more likely to have extra pulmonary TB (EPTE) and mortality was higher among those who had TB or history of TB. These findings are consistent with previous studies on high risk of deaths associated with HIV co-infection and other co-morbidities [44, 45, 46, 47]. In order for those who are TB-HIV co-infected to benefit from interventions such as co-trimoxazole preventive therapy (CPT) and HAART as advocated by [49], there must be HIV testing among all TB patients and intensified case findings for TB among people who are lining with HIV. Efforts in this regards need to be strengthened in Limpopo programme in order for the province to meet the global targets for all TB patients tested for HIV, and all TB patients living with HIV provided with an ART, and isoniazid preventive therapy for HIV positive people without active TB [49].

Clinics which are a primary health facility, are available, accessible and provides ARV treatment to all HIV/AIDS patients in Limpopo. However, distance to clinics and transportation remain a problem in rural areas of Limpopo province. Secondly, these rural based clinics close early due to staff shortage and high crime levels, which make it dangerous for staff to work late. The findings of this sudy proves that these clinics provide good services to HIV/AIDS patients resulting in low mortality rate. Only the specialised site 2.3c, Table 3.8 and Table 2.3 have highest mortality rate. The patients in these health facilities (Specialised Sites) are referred HIV/AIDS patients from the clinics.

SOFTWARE

The SAS system, version 9.4, was used to perform all of the analyses and graphical presentation in this study. R version 3.5.0 (2018-04-23) , the Coxph(Surv(.)) procedure used to fit accelerated time models, Coxph was used to fit proportional hazards models, and survfit(Surv(.)) was used to compute Kaplan-Meier estimates. Furthermore, Proc Lifetest was used to in the computation of the residuals plotted in the exponential, Weibull, Log-normal, Log-logistic and generalised gamma models described in the sub-section ”Model diagnostics”.

2.8 Study Limitation

The study limitation was that there were no sufficient statistical power due to minimum death rates (30%) with a very high censored (95.12%) HIV/AIDS patients in Limpopo Province. Secondly, the secondary data did not capture individuals level of education of which according to literature plays a vital role in the spread of HIV. Thirdly, the demography of individuals were not also considered which in one’s opinion would improve the data analysis. Fourthly, Ethical groups were also not considered when individual profile data captured. Due to big data, the researcher would not not able to group HIV/AIDS patients into groups,to establish which group of Limpopo population has the highest HIV/AIDS incidence.

The interpretation of these results is subject to many important limitations. There were a substantial amount of missing data in this analysis, for example, there was high proportion of missing values for WHO clinical stages, and viral load. Patient’s occupational information was not available in Limpopo Province dataset to reflect prevalence among South African health workers, namely, medical professional and non-medical professional health workers. Research done by Shisana *et al.*, (2004) in health workers across South Africa found that non-professional had a HIV prevalence of 20.3% while professional had a prevalence of 13.7%. It is unknown as to whether similar cases in Limpopo as a Province do exists or not.

The researcher lacked reliable CD4 cells counts, which are established predictors of morbidity and mortality in our present study. However, the results strongly suggest that simple CD4 counts measurements can be useful alternative prognostic marker. There might have been biased reporting, although the direction of such bias is difficult to predict.

Chapter 2 of this study managed to address the following objectives: it compared the average evolutions between gender, districts, health care facilities, previous opportunistic infections, and AIDS clinical stages using Kaplan-Meier curves; it again compared the semi-parametric and parametric models; analysed survival data using both Cox proportional hazard and parametric hazard models.

2.9 Conclusion

In this study, the Cox model and Accelerated Failure Time(AFT) model have been compared using HIV/AIDS patients data of Limpopo Province. The AFT model was fitted and diagnosed using Cox-Snell residuals, and the generalised gamma model provided a better fit to the studied Limpopo HIV/AIDS data with a lowest AIC value as a result the generalised gamma is a preferred model. The results obtained from Kaplan-Meier curves show that males survival time were shorter than their counterpart, and Mopani district experienced more deaths due to HIV infections than the other four districts. Cox proportional hazard models were better than AFT based on sign of coefficients.

We believe that our results better reflect the reality in a rural, peri-urban and urban of Limpopo Province, and thus, may be applicable to other provinces of South Africa with similar settings. Stigma and delay in seeking health care, lack of voluntary testing and counseling services, and health system delays in referral and ART initiation are possible reasons for continued progression to advanced stage of HIV/AIDS. The findings also indicated that it is critical that HIV/AIDS patients are diagnosed earlier in the clinics and referred to start with antiretroviral treatment. This sudy, Mopani district stood out to be a Limpopo District with the highest patients with HIV prevalence due to the fact that Mopani has more villages. In many intervention, a special attention should be given to people with a high risk of infection, which include but not limited to commercial sex workers but also migrant and partners of migrants. The TB prophylaxis drug could substantially reduce TB morbidity and mortality among those with HIV and this is particularly important in the context of copper mines in Phalaborwa, platinum mine Mokopane, Burgerfort, Thabazimbi and Northam as well as chrome in Burgerfort, where the high rate of silicosis and HIV lead to a situation in which the incidences of TB is about 3000 per 100 000 men per year [83]. More importantly, there has been few intervention programmes by provincial government, even on small scale, which attempt to reduce transmission among migrants and their rural or peri-urban or urban partners. Policy issues need to be addressed, including the nature and extent of migration, the rights of migrants, and the kind of services to which they have access. That must be conducted for those in both the formal and informal sectors, and even illegal migrants must be able to access the health services without fear of exposure. Finally, unless the issues of migration and disease are well understood and dealt with effectively, it unlikely that greater battle to control and manage AIDS will be won. The ultimate solution to the problem of pediatric AIDS lies in prevention of mother-to-child-transmission (MTCT) and prevent primary infection to women. A population-based household survey in

South Africa found that 42% of those with HIV infection in South Africa were men [84]. A challenge is still that additional effort are needed to attract and retain men on treatment.

2.10 Recommendation

The effective HIV prevention remains an urgent priorities in South African strategies to date may have been partly effective in reducing risk among educated and mobile members of the society. It is possible that changes will emerge in all groups over time as safer-sex behaviour diffuse, leading to reduction of HIV prevalence as witness to other Sub-Saharan African countries. More HIV/AIDS education and awareness to rural people is recommended so that they can stop consulting with traditional healers, because by the time they are diagnosed, they are either in symptomatic stage or advanced stage of HIV where kidneys begin to fail. If government can create more job in rural areas that will results in reduction of unemployment and migrant labours who leave their families for months or so and work in urban areas or mines. Secondly,there should be structural development that would be either bring labour markets closer to the rural setting or migrants frequently return to families. High level of knowledge and positive attitudes towards HIV prevention is recommended.

Chapter 3

Modelling of HIV/AIDS Dataset using Linear Mixed-Effects Models

3.1 Introduction

The linear mixed model is a standard statistical methods to analyse change of time of a longitudinal Gaussian outcome and assess the effects of covariates on it [134, 126, 175, 117]. The earliest methods for analysis of longitudinal data was a mixed-effects analysis of variance (ANOVA), with a single random subject effects. The inclusion of a random subjects effects induced positive correlation among the repeated measurements on the same subjects.

There exist different methods for analysing longitudinal data, and they mostly based on generalised linear models (GLM), however, GLM have their own statistical weaknesses, because they violate the assumption of independent observations. GLM estimate the model accurately on condition that data is of repeated-measure and balanced design. Unfortunately, that is not always true in practice, and that condition (unbalanced repeated design) is hard to meet and the use of traditional univariate and multivariate test statistics might increase Type 1 error under this condition [94, 106, 107]. In most cases observations may not truly be independent due to high-level of clustering unit (consequently, clustering unit are correlated). The violation of independence observations in longitudinal data do not pose a serious problem to researchers because there exist a statistical technique to circumvent that problem whenever it exists [107, 108, 109]. There is an increase interest to study the rate of change using individual growth curve (IGC) models. The IGC is an advanced technique for modeling within-person systematic change and between-person differences in developing outcome across different measurement of viral load over time. After we had specified different sets of models, we were able to examine change in predictive effective way after addition of covariates. Many researchers advocate for the use of IGC when examining the longitudinal pattern over time [106, 110].

The advantage of this technique provides researchers with more flexible and powerful approach when handling an unbalanced data(inconsistent time interval, missing data, unequal sample size). Secondly, it allows researchers to study both intra- and inter-individual differences in the growth parameters (e.g., slope and intercept) [111]. Thirdly, IGC retains all of the information and variability in the data when examining the rate of change in the independent variables [113]. That information was valuable in this study because it captured not only individuals variations in their initial status, but also in their rate of changes. Fourthly, IGC analyses estimate the changes parameters with greater precision when number of viral load increased. Consequently, it improves the reliability of the growth parameters by reducing standard errors of the within-subjects change in growth parameters estimates [110, 113]. Clearly, this has an enormous advantage as compared to GLM. Fifthly, the effects of predictors at higher-level (e.g., type of facilities, districts) and other predictors on the individual growth can flexibly be added in the growth curve models [114]. In addition, it allows predictors of growth to be discrete or continuous as well as time-variant or time-invariant. The time-variables in this study are gender and age (age at baseline), type of health care facilities, districts and previous opportunistic infections.

Lastly, IGC is more powerful than other methods in examining the effects associated with repeated measurements as it model's the covariance matrix, rather than imposing a certain type of structures as commonly used in traditional univariate and multivariate approach [115]. The covariance structure of repeated measurements can be specified in IGC models, and, it allow researchers to examine true change and possible determinants of the structures by choosing an appropriate covariance structure for growth curve model, the variance would be reduced and allow researchers to specify a correct model that conceptualises the patterns change over time. The alternative to GLM, the linear mixed-effects models are commonly used for analyses of unbalanced, repeated-measurements design to understand change rate over time. Mixed-Effects models as presented by Harville (1997), Laird and Ware (1982), Jennrich and Schluchter (1986), Bates *et al.*, (2014), Verbeke and Molenberghs (2000), Fox (2016) and others, have become popular for the analysis of longitudinal data because they are flexible and widely applicable. They are commonly used by various fields of social sciences, medical and biological sciences. In mixed effects models, it is assumed that the unobserved heterogeneity at cluster levels cause intra-cluster correlation between responses, and hence the mean level of the responses and which can vary across clusters. Fixed effects and random effects are used to model such intra-cluster correlation. The main difference between fixed effects and random effects is that fixed effects assume that unobserved heterogeneity at cluster level is constant while random effects assume that such quantity is random. Thus, the estimation of fixed effects concern the actual sizes of the cluster specific effects. When the number of clusters become large the number of fixed effects coefficients increases rapidly. However, the researchers are more interested in the distribution of random effects rather than the actual size of random effect coefficients. Generally, the random effects assumed to follow a zero mean multivariate normal distribution, and its covariance matrix become our key interest because it summarises the intra-cluster correlation. The challenge becomes when the number of random components becomes large, then the estimation of random effects in a mixed effects model involves a high dimensional covariance matrix that can greatly increase computational instability. Hence, identification of effective components of random effects is very crucial for the applied researchers to build more interpretable and ease the computational burden [125].

A general linear mixed effects (LME) models can be written as [134]:

$$\begin{aligned} \mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \boldsymbol{\varepsilon}_i,\ i = 1, 2, .., n. \\ \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \\ \varepsilon \sim N(\mathbf{0}, \Sigma_i), \end{aligned} \tag{3.1}$$

where β_i=($\beta_1, \beta_2, ..., \beta_p)^t$ is a $p \times 1$ vector of fixed effects, $\mathbf{b_i} = (b_{i1}, b_{i2}, ..., b_{iq})^t$ is a $q \times 1$ vector of random effects, the $n_i \times p$ matrix X_i and the $n_i \times q$ matrix Z_i are known design matrices may contain covariates, ε_i=($\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})^t$ represents random errors of the repeated measurements within-individual i (cluster), D is a $q \times q$ variance-covariance matrix of the random effects, and Σ_i is a $n_i \times n_i$ covariance matrix of the within-individual errors. We assume that $\Sigma_i = \sigma^2 I_{n_i}$ (homoscedastic conditional independence model) where I_{n_i} is the $n_i \times n_i$ identity matrix, i.e., the within-individual measurements are assumed to be independent with constant variance. The value of σ^2 represents the magnitude of the individual variation, and the value of Σ represents the magnitude of the between-individual variation, Wang and Taylor (2000) showed that LME model 3.1 is always identifiable if $\Sigma_i = \sigma^2 I_n$, The fixed effects β are population-level parameters and are the same for all individuals, while random effects b_i are individual-level, representing individual variation from population-level parameters. Since individual shares the same random effects, the multiple measurements within each individual or cluster are correlated. The linear mixed effects (LME) model allows unbalanced data in the response which is an advantage of mixed models.

It follows from (3.1) that $\mathbf{Y_i} \sim N(X_i\beta_i + Z_ib_i, \Sigma_i)$ conditional on random effect b_i. Let us suppose that $f(\mathbf{y}_i|\mathbf{b}_i)$ is a conditional density function of $\mathbf{Y}_i$ and $f(\mathbf{b}_i)$ be the corresponding density function. Thus the marginal density function of $\mathbf{Y}_i$ is given by

$$f(\mathbf{y}_i) = \quad f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)\mathbf{db}_i \tag{3.2}$$

where $f(\mathbf{y}_i|\mathbf{b}_i) \sim N(X_i\beta, Z_iDZ_i^t + \sigma^2 I_n)$ and $f(\mathbf{b}) \sim N(0, D)$ which can be shown to be the density function of a n_i-dimensional normal distribution with mean vector $X_i\beta$ and variance-covariance matrix $V_i = Z_iDZ_i^t + \Sigma_i$.
The distributional assumption is made for effects in (3.1), since the sampled subjects are thought to represent a population of subjects. The matrix D is usually unstructured, but it can be structured such as a diagonal matrix [88].

Covariance Structure choices

Let us assume covariance structure is given by $\Sigma_i = Z_i\Sigma Z_i^t + R_i$ which depends on Σ, and R_i is $n_i \times n_i$ matrices. We can choose some structure for Σ in the following possible way:

- unstructured: all $(q+1)(q+2)/2$ unique parameters of Σ are free.

- variance components: σ_k^2 free and $\sigma_{kl} = 0$ if $k /= l$

- Compound symmetry : $\sigma_k^2 = \sigma_k^2 + \sigma^2$ and $\sigma_{kl} = \sigma_v^2$
- Autoregressive (1): $\sigma_{kl} = \sigma^2\rho^{|k-1|}$ where ρ is autocorrelation.
- Toeplitz: $\sigma_{kl}=\sigma^2\rho^{|k-1|} + 1$ where $\rho_1 = 1$

Unstructured Covariance Matrix

Let consider unstructured covariance matrix where all $(q + 1)(q + 2)/2$ unique parameters of Σ are free. For example with $q = 3$ we have $b_i = (b_{i0}, b_{i1}, b_{i2}, b_{i3})$ and

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \sigma_{01}^2 & \sigma_{02}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_1^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{20}^2 & \sigma_{21}^2 & \sigma_2^2 & \sigma_{23}^2 \\ \sigma_{30}^2 & \sigma_{31}^2 & \sigma_{32}^2 & \sigma_3^2 \end{pmatrix}$$

where 10 free parameters are the 4 variance parameters $\{\sigma_k^2\}_{k=0}^3$ and the 6 covariance parameters $\{\sigma_{kl}\}$, $b_i = (b_{i0}, b_{i1}, b_{i2}, b_{i3})$

Parameter Estimation for fixed effects

Let α denote the vector of all variance-covariance parameters found in $V_i = Z_iDZ_i^t + \Sigma_i$ (α consists of the q(q+1)/2 different elements in D and all parameters in Σ_i). Again, let $\theta = (\beta^t, \alpha^t)^t$ be the vector of all parameters in marginal model $\mathbf{Y}_i \sim N(X_i\beta, Z_iDZ_i^t + \Sigma_i)$.

Thus, the marginal likelihood function is given by

$$L(\theta) = \prod_{i=1}^{N} \{(2\pi)^{-\frac{n_i}{2}}|V_i(\alpha)|^{-\frac{1}{2}} \times \exp(-\frac{1}{2}(\mathbf{Y}_i - X_i\beta)^t V_i^{-1}(\alpha)(\mathbf{Y}_i - X_i\beta))\} \qquad (3.3)$$

Case 1: Assume that α is known, then the maximum likelihood estimator (MLE) of β, conditional on α is given [134] by

$$\hat{\beta} = \sum_{i=1}^{N} (X^t W_i X_i)^{-1} \sum_{i=1}^{N} X^t W_i \mathbf{y}_i \tag{3.4}$$

and its variance-covariance is given by

$$var(\hat{\beta}) = \sum_{i=1}^{N} (X^t W_i X_i)^{-1} \tag{3.5}$$

where $W_i = V_i^{-1}(\alpha)$.

A sufficient condition for (3.4) to be unbiased it is that the E(Y_i) should be correctly specified as $X_i\beta$, as well as $var(Y_i) = Z_i D Z_i^t + \Sigma_i$. Liang and Zeger (1986) proposed inferential procedure based on sandwich estimator for var($\hat{\beta}$) by replacing var(Y_i) by $\mathbf{r}_i\mathbf{r}_i^t$, where $\mathbf{r}_i = \mathbf{y}_i - X_i\hat{\beta}$, and the resulting estimator found to be consistent as long as E(Y_i) was correctly specified.

Case 2: When α is unknown.

Let us assume that the estimator of α is available, then we can set $\bar{V}_i = V_i(\hat{\alpha}) = \bar{W}_i^{-1}$. We can now replace W_i in (3.4) by $\bar{W}_i$. The standard error of β can be obtained by replacing α by $\hat{\alpha}$. Dempster *et al.*, (1981) found that this approach has weakness in the sense that it does not take into account the variability introduced by estimating α that leads to the underestimation of the variability of $\hat{\beta}$. Dempster et al., (1977) proposed EM algorithm for the calculation of MLE's based on incomplete data and how it can be used for estimation of variance components in mixed-model analysis of variance. A decade later, Lindstrom and Bates (1988) used Newton-Raphson-based procedures to estimate all parameters $\theta = (\beta^t,\ \alpha^t)^t$ in the mixed-model. The Statistical inference for a linear mixed effects (LME) model is based on maximum likelihood (ML) method or restricted maximum likelihood (REML) method [134, 90].

REML is by default in many softwares, but we need to use maximum likelihood if we want to conduct likelihood ratio test. For the purpose of this study, both fixed effects β and random effects b are estimated as follows

$$\hat{\beta} = \sum_{i=1}^{n}(X_i^t V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i^t V_i^{-1} y_i,$$

where $V_i = Z_iDZ_i^t + \sigma^2 I_{n_i}$ and $\hat{b}_i = \hat{D}\Sigma_i^t\Sigma_i^{-1}(y_i - X\hat{\beta})$.

Parameter Estimation for random effects

In practice researchers are more interested in estimating the parameters in the marginal linear mixed models (i.e., the fixed effects β and the variance components D and σ^2). However, it is quite useful and helpful to estimate the random effects $\mathbf{b}_i$, which reflects how much subject-specific profiles deviate from the overall mean profile evolving differently in time. It is also helpful to detect special profiles(i.e., individuals outliers) [172]. For the purpose of estimating random effects $\mathbf{b}_i$, we will use hierarchical model (3.3) as appropriate model, because the variability in the data can be explained these random effects. The random effects naturally represents heterogeneity between subjects variability in the population, and that is also the case with our Limpopo AIDS dataset. Since the random effects in model (3.3) is assumed to be random variables, Box and Tiao (1992), and Gelman *et al* (1995), used Bayesian approach to estimate $\mathbf{b}_i$.

The marginal distribution of $\mathbf{b}_i$ is a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance D. In Bayesian literature, N($\mathbf{0}$,D) is called prior distribution of the parameter $\mathbf{b}_i$ because it does not depend on $\mathbf{Y}_i$.

Let us suppose that $f(y_i|b_i)$ is density function of Y_i conditioned on b_i , and $f(b_i)$

be a prior density function of b_i, and thus the posterior density of b_i is given by:

$$f(\mathbf{b}_i|\mathbf{y}_i) = \frac{f(\mathbf{y}_i|\mathbf{b}_i)}{\int_{\mathbf{b}} f(\mathbf{y}_i)|\mathbf{b}_i f(\mathbf{b}_i)d\mathbf{b}_i} \tag{3.6}$$

Lindley (1972) and Smith (1973) used the theory on general Bayesian linear models and showed that $f(\mathbf{b}_i)$ is estimated by the mean of posterior distribution in (3.6) given by:

$$\begin{aligned}\hat{\mathbf{b}}_i(\theta) &= E[\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i] \\ &= \mathbf{b}_i f(\mathbf{b}_i|\mathbf{y}_i)d\mathbf{b}_i \\ &= DZ_i^t W_i(\alpha)(\mathbf{y}_i - X_i\beta)\end{aligned} \tag{3.7}$$

and the covariance matrix of corresponding estimator is;

$$var(\hat{\mathbf{b}}_i) = DZ_i^t\{W_i - W_iX_i(\sum_{i=1}^{N} X_i^tW_iX_i)^{-1}X_i^tW_i\}Z_iD, \tag{3.8}$$

where $W_i = V_i^{-1}$ [134]. Since the var($\mathbf{b}_i$) underestimate the variability in ($\hat{\mathbf{b}}_i - \mathbf{b}_i$), we usually uses

$$var(\hat{\mathbf{b}}_i - \mathbf{b}_i) = D - var(\hat{\mathbf{b}}_i) \tag{3.9}$$

to assess the variation in ($\hat{\mathbf{b}}_i - \mathbf{b}_i$), [134].

The unknown parameter β and α in (3.7), (3.8) and (3.9) are replaced by their maximum likelihood or restricted maximum likelihood estimates, and the resulting estimates for $\mathbf{b}_i$ are called Empirical Bayes estimates.

Other estimation methods of linear mixed models

Likelihood Ratio Test

Given two nested models (M_1 and M_0) the likelihood ratio test (LRT) is given by [116]:

$$D = -2 ln \frac{L(M_0)}{L(M_1)} = 2[LL(M_1) - LL(M_0)]$$

where

- L(·) and LL(·) are the likelihood and log-likelihood, respectively.
- M_0 is null model with p parameters.
- M_1 is an alternative model with p+k parameters.

Wilks's theorem reveals that as $n \to \infty$ (the sample become very large) then $D \sim x_k^2$ where x_k^2 denotes chi-squared distribution with k degree of freedom.

Inference for Fixed effects

We will use LRT idea to test fixed effects as follows:

$$H_0 : \beta_k = 0 \quad versus \quad H_1 : \beta_k \ /= 0$$

and then compare D with x_k^2

Inference for random effects

$$H_0 : \sigma_{jk} = 0$$

versus

□ $H_1 : \sigma_{jk} > 0$ if $j = k$

□ $H_1 : \sigma_{jk} \ /= 0$ if $j \ /= k$.

where σ_{jk} denotes the entry in cell j, k of Σ. We will use LRT idea to test hypotheses and compare to [116]:

- x_k^2 distribution if j /= k
- mixture of x_k^2 and 0 if j=k

3.2 Aim

The aim of this study was to apply linear mixed-effects models techniques in the analysis of HIV/AIDS patients in Limpopo Province, South Africa.

3.2.1 Objectives:

In this chapter we will be addressing the following objectives:

i) to describe the relationship between response variable and the covariates using linear mixed effect models;

ii) to show how longitudinal evolution of viral load is associated with time-to-death;

iii) to characterise viral dynamics in patient population and intra- and inter-subject variation;

iv) to assume random effects that gives some structure to error terms that characterises individual variation due to some factor levels; and

v) to demonstrate non-linear statistical framework as a basis for estimation of population and individual viral dynamics parameters and how models may be used to draw biological relevant interpretations and aid clinical decision-making within the context of Limpopo HIV/AIDS data.

3.3 Dataset

In this study, we used secondary data obtained from Limpopo Department of Health for repeated viral loads measurements of HIV/AIDS patients for the period January 2011 to 2016 January. After cleaning data, 6439 (69.9%) c patients were censored and 2776 (30.1%) patients had died. The following variables were recorded: gender, AIDS clinical stage as stipulated by WHO, district, previous opportunistic infection (e.g., tuberculosis), event (i.e., died), type of health care facilities, CD4 count cell (at baseline) and viral loads. The viral loads were recorded at 3,6,12, 24,..., 132 months after initiation of ART treatment. The baseline CD4 cell counts were transformed by using square root, and viral load by natural logarithmic, and, the transformation of these raw data (viral loads and CD4 cell counts) were necessary in order to stabilise their variances(it is more normally distributed). The viral load constitute an important marker of the strength of immune system. Hence, when the viral load of patient decreases it is an indicative that the condition of the immune system of patient improves.

The study protocol was approved by the Turfloop Research Ethics Committee (TREC) of University of Limpopo. In addition, permission was obtained from the Limpopo Province Department of Health Research Committee to use their secondary data. All data captured were without specific patient identifiers, to ensure the anonymity of the patients, and all the information obtained was treated with utmost confidentiality.

3.4 Methodology

Let $\mathbf{Y}_{ij} = (y_{i1}, y_{i2}, ..., y_{in_i})^t$ be the n_i repeated measurements of the response variable Y of patient i, $i = 1, 2, \ldots, n$, at time t_{ij} ,$j = (1, 2, \ldots, n_i)$.

Random Intercept Models with Linear Time.

A random intercept model with linear time has the following form:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + \varepsilon_{ij} \tag{3.10}$$

where

- β_0 is fixed effect intercept of the model.
- β_1 is the fixed slope of the model.
- t_{ij} is the time variable for the j-th measurement of the i-th patient.
- $b_i \sim^{iid} N(0, \sigma_b^2)$ is the random intercept for the i-th patient.
- $\varepsilon_i \sim^{iid} N(0, \sigma_\varepsilon^2)$ is the Gaussian error term.

The random intercept model's assumptions are:

- the relationship between T and Y is linear.
- y_{ij} and t_{ij} are observed random variables.
- $b_i \sim^{iid} N(0, \sigma_b^2)$ is an unobserved random variable.
- $\varepsilon_i \sim^{iid} N(0, \sigma_\varepsilon^2)$ is an unobserved random variable.
- b_i and ε_i are independent of one another.
- β_0 and β_1 are unknown constants.
- $(y_{ij}|t_{ij}) \sim N(\beta_0 + \beta_1 t_{ij}, \sigma_Y^2)$ where $\sigma_Y^2 = \sigma_b^2 + \sigma_\varepsilon^2$

Furthermore, we assume the covariance structure for random intercept model as follows: The conditional covariance between any two observations is

$$Cov(y_{hj}, y_{ik}) = \begin{cases} \sigma_b^2 = \omega\sigma_y^2, & \text{if } h = i \quad and, j = k, \\ 0, & \text{if } h \neq i. \end{cases}$$

where $\omega = \frac{\sigma_b^2}{\sigma_Y^2}$ is the correlation between any two repeated measurements from the same patient.

- if $h = i$, then $Cov(y_{ij}, y_{ik}) = E[(b_i + \varepsilon_{ij})(b_i + \varepsilon_{ik})] = \sigma_b^2$
- if $h /= i$, then $Cov(y_{hj}, y_{ik}) = E[(b_h + \varepsilon_{hj})(b_i + \varepsilon_{ik})] = 0$

Random Intercept and Slope Models with Linear Time

A random intercept and slope model with linear time has the following form:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij} \tag{3.11}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ where

- y_{ij} is the response for j-th measurements of i-th subject.
- β_0 is fixed effect intercept of the model.
- β_1 is the fixed slope of the model.
- β_{i1} is the linear time fixed slope for the i-th subject.
- t_{ij} is the time variable for the j-th measurement of the i-th subject.
- $b_i \sim^{iid} N(0, \sigma_b^2)$ is the random intercept for the i-th subject.
- $\varepsilon_i \sim^{iid} N(0, \sigma_\varepsilon^2)$ is the Gaussian error term.

The fundamental assumptions of the random intercept and slope models are:

- the relationship between T and Y is linear;
- t_{ij} and y_{ij} are observed random variables;
- $b_{i0} \sim^{iid} N(0, \sigma_{b_0}^2)$ and $b_{i1} \sim^{iid} N(0, \sigma_{b_1}^2)$ are unobserved random variable;

- $(b_{i0}, b_{i1}) \sim^{iid} N(0, \Sigma)$, where:

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$$

- $\varepsilon_{ij} \sim^{iid} N(0, \sigma_e^2)$ is an unobserved random variable;
- (b_{i0}, b_{i1}) and ε_{ij} are independent of one another;
- b_0 and b_1 are unknown constants; and
- $(y_{ij}|t_{ij}) \sim N(b_0 + b_{ij}, \sigma_{ij}^2)$ where $\sigma_{ij}^2 = \sigma_0^2 + 2\sigma_{01}t_{ij} + \sigma_1^2 t_{ij}^2 + \sigma_\varepsilon^2$

3.5 Quadratic time model using natural splines

Quadratics splines

A random intercept and slope model with non-linear time has the following form:

$$y_{ij} = \beta_0 + \beta_{i1}t_{ij} + \beta_{i2}t_{ij}^2 + b_{i0} + b_{i1}t_{ij} + b_{i2}t_{ij}^2 + \varepsilon_{ij} \tag{3.12}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ where

- y_{ij} is the response for j-th measurements of i-th subject.
- β_0 is fixed effect intercept of the model.
- β_1 is the fixed slope of the model.
- β_{i1} is the linear time fixed slope for the i-th subject.
- β_{i2} is the quadratic time fixed slope for the i-th subject.
- t_{ij} is the time variable for the j-th measurement of the i-th subject.
- t_{ij}^2 is the quadratic time variable for the j-th measurement of the i-th subject

- $b_i \sim^{iid} N(0, \sigma_b^2)$ is the random intercept for the i-th subject.
- $\varepsilon_i \sim^{iid} N(0, \sigma_\varepsilon^2)$ is the Gaussian error term

Clearly, b_{i0} allows each subject to have unique intercept, and b_{i1} allows each subject to have unique slope.

3.6 Statistical Analysis

3.7 Results

The data were analysed using mixed effects models with maximum likelihood estimation. This method modeled individual change over time, explore differences in change, and examine the effects of covariates and rate of growth.

Model 1: Unconditional Linear Growth Curve Models

The unconditional linear growth model is given by

$$Y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \varepsilon_{ij}$$

where

- β_0 is the initial status for individual i.
- β_1 is the linear rate of change for individual i.
- t_{ij} is the linear time for individual at time t.
- ε_{ij} is the residual in the outcome variable for individual i at time time t.

The significant values in both the intercept and linear slope parameters indicate that the initial status and linear growth rate were not constant over time. The was significant linear increase in viral load ($\beta = 0.00154$, SE=0.00076,

Table 3.1: Information Criteria

Log Likelihood	-21979.46
Akaike's Information Criterion	43966.91
Bayesian Information Criterion	43997.6

Table 3.2: Estimates of fixed effects

Parameter	Estimates	Std. Error	P-value
Intercept	2.176699	0.00962	< 0.05
Time	0.001537	0.00076	< 0.05

$P < 0.05$). The mean estimated initial status and linear growth rate for the sample were 2.18 and 0.0015, respectively. That suggest that viral load was 2.18 and 0.002 increased with time. The random error terms associated with the intercept and linear time were significant ($P < 0.05$), suggesting that the variability in these parameters could be explained by between-individual predictors. The correlation ($\beta = -0.502$) between intercept and linear growth parameter was negative. This suggest that patients with high viral loads had a slower linear decrease, whereas patients with low viral load had a faster decrease in the linear growth over time.

Model 2: quadratic growth curve models

The quadratic growth curve model is as follows

$$Y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j}(t_{ij}^2) + \varepsilon_{ij}$$

where

- β_0 is the initial status for individual i.
- β_1 is the linear rate of change for individual i.
- t_{ij} is the linear time for individual at time t.
- ε_{ij} is the residual in the outcome variable for individual i at time time t.

Table 3.3: Estimates of covariance parameters

Residual	0.662977
Intercept variance	0.54504
Correlation	-0.50200

Table 3.4: Information Criteria

Log Likelihood	-21819.22
Akaike's Information Criterion	43652.45
Bayesian Information Criterion	43706.16

The results in Table 3.5 shows that all growth parameters were significant ($P < 0.05$) indicating that the were significant between-patients variations in the initial status, and linear and quadratic time trajectories. The significant linear effects for viral load was positive ($\beta = 0.4500$, SE=0.0906,$P < 0.05$) revealing that the rate of linear growth increase over time. The quadratic effect was also positive ($\beta = 1.6098$, $SE = 0.2334$,$P < 0.05$), showing that the rate of growth increase over time. Compared to the linear trajectory (0.45), the rate of quadratic was (1.6098) greater. Based on the above results, it showed that viral load marker increased from the beginning, and the trend continued.

3.7.1 Model selection

The quadratic model improved model fit over the linear model, hence both linear and quadratic curve parameters were retained. Thus, indicate that the potential of curvature trajectories fit the data better.

3.7.2 Model diagnostics

Figure 3.1 shows Q-Q plots, and the points seem to fall about the straight line. The x-axis plots are the quantiles from the standard normal distribution with mean zero(0) and variance one(1). The quantiles plot does not raise any significant concern with the normality of the weighted residuals.

Table 3.5: Estimates of fixed effects

Parameter	Estimates	Std. Error	P-value
Intercept	2.19827	0.01040	< 0.05
$Time$ effect	0.45005	0.090618	< 0.05
$(Time)^2$ effect	1.609814	0.23345	< 0.05

Table 3.6: Model selection

Model	AIC	BIC	Log-Likelihood	L.Ration	P-value
Linear Model	44231.12	44261.81	-22111.56		
Quadratic Model	44175.15	44228.87	-22080.58	61.96343	0.0001

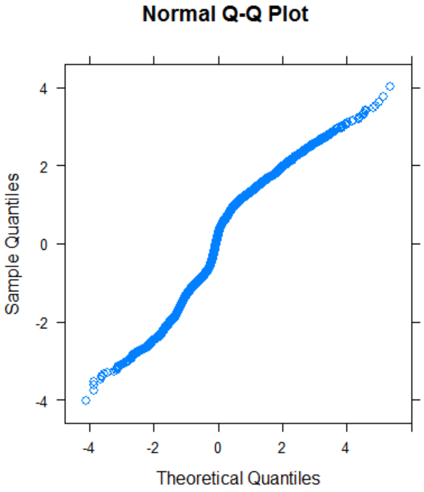


Figure 3.1: Normal Q-Q-plot

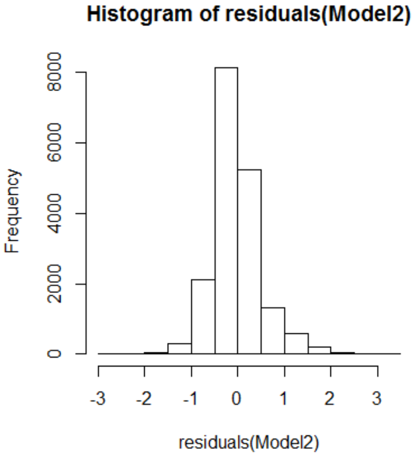


Figure 3.2: Weibull AFT Model

Model 3: Adding covariates to quadratic growth curve models

$$y_{ij} = \gamma_{0i} + \gamma_{1i} time + \gamma_{2i}(time)^2 + \gamma_{ji} W_j + b_{ji} + \varepsilon_{ji}$$

where

- y_{ij} is the grand mean for viral load for the whole sample at time t.
- γ_{0i} is the initial status of viral for the whole sample at time t.
- γ_{2i} is the quadratic slope of change relating to viral load for the whole sample at time r.

- b_{ji} is the random effects(amount of variance) that are unexplained by the covariates.
- γ_{ji} is used to test whether the covariate is associated with growth parameter.
- W_j is an explanatory variable to predict on inter-individual variation on outcome variable.
- ε_{ij} is an error term assumed to be independent and normally distributed, and the variance is equal across individual.

Table 3.7: Information Criteria

Log Likelihood	-21229.24
Akaike's Information Criterion	42536.49
Bayesian Information Criterion	42835.66

Table 3.8 of estimate of fixed and random effects model, gender, age at baseline, CD4 counts, Sekhukhune district, Waterberg district, Clinical stages, community health care (CHC), District hospitals, Non-Medical hospitals were significant predictors of linear and quadratic models for patient's viral load, but not associated with initial status, while Mopani district, Vhembe district, previous opportunistic infection(e.g. TB), Provincial Tertiary Hospital, Regional hospitals, Specialised Psychiatric hospitals were not. Regarding linear slope of patients viral load, the male patients showed moderately faster rate of change of viral loads as compared with female patients($\beta = 0.009$). In terms of quadratic growth, the male patients had faster rate of change as compared with female patients. That suggest that the immune system of female patients is better than that male patients. Patient's both linear and quadratic slopes age showed negative rate of change, which indicate that viral load growth rate decrease with age. Thus, patient's immune system improves as he grows older. The patient linear and quadratic growth rate of CD4 counts is decreasing

(($\beta = -0.0110$),($\beta = -0.0867$), respectively) which indicate that patient's viral load change rate is increasing faster. For three Limpopo Provincial districts, namely, Mopani, Sekhukhune and Waterberg districts both linear and quadratic rate of change are faster as compared to Capricorn district. That suggests that the growth rate of change of patient's viral load in those districts is growing faster as compared to Capricorn district, which indicate that there were more deaths. The Vhembe district linear slope ($\beta = 0.1015$) showed positive rate of change but the quadratic slope ($\beta = -.5991$) showed negative rate of change. That suggested that viral load linear growth rate was faster and quadratic growth rate was slower as compared with Capricorn district.
When we investigated the estimates for parameters in the mean structure shows that no significant interaction seems to be present between the gender, CD4 cell counts, districts, previous opportunistic infection health care facilities effects and linear and quadratic time effects, suggesting that only a patient's intercepts are influenced by gender, CD4 cell counts, districts, previous opportunistic infections, and health care facilities and not the complete evolution of viral load over time. However, there is significant interaction between age effects and linear and quadratic time effects, which suggest that patient's intercept and age effects is influenced by age for the entire evolution of viral load over time.

Regarding the linear and quadratic slope (($\beta = 0.19725$),($\beta = 1.4838$), respectively) of patients who had experience previous opportunistic infection (e.g., tuberculosis) showed positive slopes, which indicate the growth rate of change of patient's viral load was faster as compared to patients who had no previous opportunistic infections. That suggests that immune system of patients who did not experience opportunistic infection in the initiation of ARV were better.

The predictors accounted for 1.1% ((($0.6698313 - 0.6626046) \backslash 0.6698313) = 0.0108$)

of the within-individual variations in patients viral load. That shows that only 1.1% of the overall variability in the patient's viral load is explained by patients predictors. Singer *et al.*, (2003) proposed using prototypical values to demonstrate the effect treatment on initial status and the rate of change across time, and that can be achieved by plotting graphs in regression [92]. We can obtain the fitted trajectories by substituting estimated values in quadratic model: $Y_{ij} = 3.090314 + (0.503488)Time + (4.611126)Time^2 + (0.254018)GENDER + (-0.023747)Age + (-0.011201)CD4 + (0.045234)Mopani + (-0.113935)Sekhukhune + (0.021547)Vhembe + (-0.098029)Waterberg + \ldots + (0.014187)Time \times GENDER + (0.918257)Time^2 \times GENDER + \ldots + (0.437142)Time \times PrevOI + (1.401098)Time^2 \times PrevOI$. This method was used by other researchers [93].

Examining Covariance Structure

One of the advantages of Individual Growth Curve (IGC) is the availability to specify the within-individual error covariance structure that best fits the data. The purpose of testing different error covariance matrices is to describe how the error is distributed [94]. It examines whether the error imposed on the error covariance structure of the parametric model fit well to the data [91]. This is critical when we examine unequally spaced and unbalanced data, which are commonly found in longitudinal studies. In fact the studies showed that the estimated variances of the parameter estimates are likely to be biased and inconsistent when repeated measurements are taken on the same individual across time [95, 96], and consequently affect the precision of estimating the appropriate model [97]. researchers advocated the use of this variance-covariance testing approach as it improves model predictions and statistic inferences, especially when examining random effects models [98, 99]. In our study, three types of covariance structures that were commonly examined in the previous studies were tested [94, 100, 101, 102].

Table 3.8: Estimate of fixed, interaction and random effects.

Parameter	Estimate	Standard Error	t-values	P-value
(Intercept)	3.033269	0.0460163	65.91732	0.0000
Time effect	1.370642	0.5301040	2.58561	0.0098
$Time^2$ effect	5.978602	1.3863940	4.31234	0.0000
Main Effects				
GENDERMale	0.228046	0.0217828	10.46910	0.0000
Age(baseline)	-0.022125	0.0006799	-32.54080	0.0000
CD4	-0.010817	0.0015159	-7.13576	0.0000
Mopani DM	0.046500	0.0266073	1.74765	0.0805
Sekhukhune DM	-0.120330	0.0308688	-3.89811	0.0001
Vhembe DM	0.026306	0.0351884	0.74756	0.4547
Waterberg DM	-0.106254	0.0369293	-2.87722	0.0040
Clinical Stage	0.052534	0.0092627	5.67160	0.0000
Previous Oppo. Infection	0.041710	0.0383993	1.08622	0.2774
Community Health Care	-0.107052	0.0265010	-4.03956	0.0001
District Hospitals	0.097384	0.0271548	3.58625	0.0003
Non-Medical Hospitals	-0.203920	0.0828470	-2.46140	0.0139
Provincial Tertiary Hospital	-0.034701	0.0625836	-0.55448	0.5793
Regional Hospitals	0.024863	0.0453837	0.54785	0.5838
Specialised Psychiatric Hosp.	-0.334497	0.6570908	-0.50906	0.6107
Interaction Effects				
GENDERMale× Time effect	0.009358	0.3028661	0.03090	0.9754
GENDERMale × $Time^2$ effect	0.643006	0.7811001	0.82321	0.4105
Age × Time effect	-0.011036	0.0094666	-1.16576	0.2438
Age × $Time^2$ effect	-0.086695	0.0249314	-3.47733	0.0005
CD4× Time effect	-0.044283	0.0226292	-1.95688	0.0505
CD4× $Time^2$ effect	-0.079587	0.0583185	-1.36469	0.1725
Mopani DM × Time effect	0.550689	0.3624434	1.51938	0.1288
Mopani DM × $Time^2$ effect	0.053720	0.9381064	0.05726	0.9543
Sekhukhune DM× Time effect	0.526100	0.3713780	1.41661	0.1567
Sekhukhune DM× $Time^2$ effect	0.464644	0.9430047	0.49273	0.6223
Vhembe DM× Time effect	0.101496	0.4729320	0.21461	0.8301
Vhembe DM× $Time^2$ effect	-0.599138	1.2454345	-0.48107	0.6305
Waterberg DM × Time effect	0.538655	0.9837451	0.54756	0.5841
Waterberg DM × $Time^2$ effect	1.817489	2.4020241	0.75665	0.4493
Previous Opportunistic Infection× Time effect	0.197246	0.6056174	0.32569	0.7447
Previous Opportunistic Infection× $Time^2$ effect	1.483802	1.5418729	0.96234	0.3360
Random Effects				
var(b_{0i})	0.46039			
var(b_{1i})	0.90038			
var(b_{2i})	0.00001			
Residual variance				
var(ε_{ij}) = σ^2	0.43380			

Unstructured Covariance Structure

The unstructured covariance structure model often offers the best fit and is most commonly found in longitudinal data as it is the most parsimonious,

Table 3.9: Estimates of covariance parameters

Residual	0.6612362
Intercept variance	0.54430634
Correlation	-0.503

which requires no assumption in the error structure [94]. In our Limpopo Province HIV, patients study, a quadratic model, the specified errors was of error covariance structure type. The print-out with specific results are displayed in Table 3.8 with fixed effects estimates, its standard errors, t-test for parameters, significance test for the estimated variance components. The estimated method was restricted maximum likelihood (REML) which is defaults in R programme.

Compound Symmetric Structure

In order to examine whether the variance and correlation between each pair of observations are constant across time points, a compound symmetry covariance structure was tested.

Table 3.10: Information Criteria

-2Log Likelihood	-21357.57
Akaike's Information Criterion	42777.14
Bayesian Information Criterion	43014.96

First-Order Autoregressive(AR(1)) covariance structure

In the First-Order Autoregressive (AR(1)) is assumed to be heterogeneous and the correlations between the two adjacent time points decline across measurement occasions.

Table 3.11: Information Criteria

-2Log Likelihood	-21338.17
Akaike's Information Criterion	42738.33
Bayesian Information Criterion	42976.15

Table 3.12: Results of Information Criterion among Three covariance Structure Models

Covariance Structure	-2log likelihood	AIC	BIC
Compound symmetry	-21357.57	42777.14	43014.96
First-Order Autoregressive	-21338.17	42738.33	42976.15

Comparison Between Three Covariance Structure Models

Based on Table 3.13, it is observed that the smallest values in the three fit criterion were found in the First-Order Autoregressive(AR(1)) model. This suggest that the First-Order Autoregressive model was the best model in fitting the data . The correlated errors terms and heterogeneous variance might be due to the results of unequally spaced times points of measurements. If the time points were closely spaced, the possibility of modelling correlated errors might be higher than those scheduled far apart [107]. Hence the use of variance-covariance approach will definitely improve model predictions.

Chapter 3 has successfully addressed the following objectives: the relationship between response variable and the covariates using linear mixed effect models; we showed how longitudinal evolution of viral load is associated with time-to-death; we characterised the viral load dynamics in patient population and intra- and inter-subject variation; and assumed random effects that gave some structure to error terms that characterised individual variation due to some factor levels; we demonstrated non-linear statistical framework as a basis for estimation of population and individual viral dynamics parameters and how models might be used to draw biological relevant interpretations and aid clinical decision-making within the context of Limpopo Province HIV/AIDS dataset.

3.8 Discussion

The study revealed that HIV/AIDS patients co-infected with tuberculosis (TB), the death rate is higher than patients that never experienced TB infection on initiation of ARV treatment. In similar study done in Limpopo Province it was found that there were more death caused by co-infected HIV patients and TB was quoted to be leading cause [118]. Other researchers revealed that tuberculosis was associated with an increase risk of AIDS and death [119]. Individual Growth Curve model revealed that the women were in better health status as compared to men,and that can be attributed to their different lifestyles in Limpopo Province. The Limpopo Provincial strategic plan support our findings in their Provincial Strategic Plan 2012-2016, [120]. The analysis in this study revealed that HIV prevalence was very high in Mopani, Sekhukhune and Waterberg districts that can be attributed to high unemployment, migration, and mostly rural setting. Due to these poor socio-economic conditions, A significant of labour force or economically active people leave their families in search of employment elsewhere, mainly in the urban areas, mining industries or other sectors available for less educated people. Other researchers support our finding in Limpopo Province [123, 85].

3.9 Conclusion

Tuberculosis in HIV-infected patients is associated with increase risk of AIDS and death. Our findings support the view that prolonged immune activation induced by TB leads to prolonged increase of HIV replication and consequently accelerated disease progression. Unemployment, Mining industry, migration and less educated people are one of the main cause of spread of HIV epidemic.

Chapter 4

Joint Modelling of Survival and Longitudinal Outcomes

4.1 Introduction

In HIV/AIDS studies, biomarkers such as viral load and CD4 counts are often collected repeatedly over time, in parallel to the time to an event of interest, such as death from any cause. These biomarkers are often measured with error, as a result, we need to account for measurements errors when looking at how a time-varying biomarker is associated with an event of interest. These repeated measurements are often longitudinally recorded for each subject. The longitudinal studies are often affected by drop-out(informative) such as death, intermittent missingness visit or late entry. The longitudinal and survival process can be simultaneously linked in joint models [160]. The longitudinal data, such as viral load can be an important predictor to time-to-event. Classical models such as the linear mixed models for longitudinal and survival data do not take into account the dependencies between these two data types. Joint modelling is a powerful method that takes into account the dependency and association between longitudinal data and time-to-event data. Thus, joint models for longitudinal data and time-to-event data are models that brings these two data types simultaneously into a single model so that one can infer dependence and association between the longitudinal biomarker and time-to-event to better understand and assess the effect on a treatment. In recent years, joint models have gain popularities amongst researchers because it reduces bias in estimates of treatment effects and provides improvement of efficiency in the assessment of treatment effects and other prognostic factors. These properties were recently demonstrated in Chen *et al.*, (2004) as well as the analysis of Eastern Cooperative Oncology Group trial $E1193$ and simulation studies given there. A less biased estimate lead to a more accurate estimate of the treatment effects. For example, if a drug reduces the hazard of a particular disease by 30%, then a joint model may lead to an estimated hazard ratio of 75%, whereas a conventional Cox model that does not incorporate the longitudinal data into analysis may yield a hazard ratio of 80%. In this particular case we conclude that joint model is less biased than Cox model. Secondly, joint models lead to estimates with smaller standard error (SE) than Cox model estimate in treatment effects. This is indeed promising because a smaller SE implies a more prices estimates. This phenomenon has a major and an important implications on the design of a study. Greater efficiency implies greater power and smaller samples sizes in designing clinical trials. Hence, incorporating longitudinal data into the design of study has the potential of yielding lower sample sizes with higher power as compared to Cox model [159].

Specifically, joint models for longitudinal data and survival data are frequently used in quality-of-life studies wherein we are interested in examining the association between patients quality-of-life and time to event end point [159]. In HIV trials, viral loads measure how much human immunodeficiency virus (HIV) is in the blood, and these measurements are often measured repeatedly, hence it is in researchers interest to examine their association with time to death.

4.2 Literature Review

In early development of joint models for longitudinal and survival data was primarily motivated from HIV/AIDS clinical trials, in particular, joint modeling of survival data and longitudinal CD4 counts. These articles included [135, 143, 136, 140, 139, 138, 142, 137, 145, 141]. Other approaches considering a multivariate longitudinal measure include [144, 146, 147]. However, an excellent overview in literature of joint modelling of longitudinal data was done by Yu *et al.*, (2004), Tsiatis and Davidian (2004), and Rizopoulos (2012).

Self and Pawitan (1992) developed joint models for a periodically observed marker of underlying disease progression and its relationship to disease-related endpoints. They used Cox Model with time-dependent covariates to specify the relationship between a marker and disease onset and use linear mixed model to describe the evolution of observed process. Partial likelihood was used in the estimation of parameters in Cox model, and assumed that variance of the covariates in linear mixed model is fixed and known. DeGruttola and Tu (1994), applied classical joint model approach, and, in the estimation of parameters made use of an adaptation of the EM algorithm where the E-step consists of finding the expected log-likelihood conditional on the observed data [185].

Faucett and Thomas (1996) modelled a continuous covariate over time and simultaneously relating the covariate to risk, and used the Markov Chain Monte Carlo (MCMC) technique of Gibbs sampling to estimate the posterior distribution of the unknown parameters of the model. Wulfsohn and Tsiatis (1997) estimated the parameters in Cox model when the longitudinal covariate is infrequently measured with measurement error, assumed linear mixed effects model for the covariate process. Estimate of parameters were obtained by maximising the joint likelihood for the covariate process and the survival process. Proust-Lima *et al.*, (2012) introduced joint latent class models (JLCM), which consider the population of subjects as heterogeneous, and assume that it consists of homogeneous latent sub-groups of subjects that share the same marker trajectory and the same risk of event. Furthermore, it assumed that a latent class structure entirely captures the correlation between the longitudinal marker trajectory and the risk of event. Due to its flexibility in modelling, the dependency between the longitudinal marker and event time, as well as its ability to include covariates JLCM is well suited for prediction problem. The method of maximum likelihood, with log-likelihood is maximised using Marquardt algorithm with strigent convergence criteria. Henderson *et al.*, (2000) proposed a linear mixed-effects model and serial correlation of longitudinal data with pure measurements error, with survival analysis was based on semi-parametric proportional hazard model with or without frailty term. The serial correlation processes allow the trend to vary with time and induce a within subject autocorrelation structure that may be thought of as arising from evolving biological fluctuations in the process about smooth trend [163]. Song *etal.*, (2002) proposed semi-parametric approach in joint modelling analysis which violated normality assumption and for the procedure that do not require parametric random effects. Since normality assumption was relaxed they proposed likelihood approach to inferences which require only that the random effects have a distribution in a possible class with smooth densities. Verbeke and Lesaffre (1997); Tao *et al.*, (1999); Heagerty and Kurland (2001); Zhang and Davidian (2001) suggested that mis-specification of random effects distribution may lead to misleading inferences on certain model parameters.

Brown and Ibrahim (2003) proposed semi-parametric Bayesian hierarchical joint models, wherein the distributional assumptions is relaxed for the longitudinal model using Dirichlet process priors on the parameters defining the longitudinal model. The resulting posterior distribution is free of parametric constrains, resulting in more robust and efficient estimates. Rizopoulos and Ghosh (2011) proposed joint model considering multiple outcomes from Bayesian approach.

4.3 Aim

The study aimed at applying various joint modelling techniques to the clustered HIV/AIDS data in Limpopo Province, South Africa, in order to come up with a good model that will simultaneously handle the survival and longitudinal outcomes.

Objective

In this chapter we will address the following objectives:

i) perform separate longitudinal and survival analyses per outcome;

ii) establish the strength of association between the longitudinal evolution of viral load and hazard rate to death;

iii) compare separate and joint models, and various association measures such as parametric joint and shared parameter models approach;

iv) Compare average evolutions between males and females;

v) show how marker-specific evolutions are related to each other (association of the evolution);

vi) compute prediction for time to death for any randomly selected HIV positive patient by considering patient's viral load; and

vii) recommend to health decision and policy-makers how the application of joint modelling techniques can be beneficiary to HIV/AIDS patients.

4.4 Methodology

4.4.1 Data collection

The secondary data used in this study were obtained from Limpopo Department of Health, South Africa. The study population consists of HIV+ patients, and started ART treatment any time between January 2011 to January 2016. Data from earlier period were excluded due to the fact that patients records were not properly kept across five districts. After data cleaning only 9215 of them satisfied inclusion criteria and hence were included in the study. At each patient visit, viral load and other covariates were recorded. Both survival and longitudinal data were extracted from patient's profiles which contained patient's identification, gender, previous opportunistic infection, districts, type of health care facilities, viral load, CD4 cell counts, age at baseline, and patient clinical stage, see Table 1.5 in Chapter 1. The viral load was transformed in order to stabilise the variance and thereby to have a more normally distributed variable. Patient's viral load is the longitudinal variable recorded at baseline at 6 , 12, 24, 36, 48, 60, 72, 84, 96, 108, 120, 132 month visits. The sample showed missing data over time due to deaths unrelated to AIDS, dropouts, missing clinic visits and transfers to other health care facilities.

4.4.2 Methods of Data Analysis

The Survival Sub-Model

The Cox proportional hazard model is the most widely used semi-parametric survival regression, particularly, when interest is on an event outcome. For this type of model, we let T^* denote the true failure time for the *i-th* subject, and C_i the censoring time, then $T_i = min(T^*, C_i)$ represents the observed failure time for *i-th* patient, $\delta_i = I(T^* \leq C_i)$ the event indicator, with $I(.)$ being the indicator function that takes the value 1 when $T^* \leq C_i$, and 0 otherwise. Now, the Cox models can be expressed as follows:

$$\begin{aligned}\lambda(t) &= \lim_{\delta t \to 0} \frac{P(t \leq T^{*} \leq t + \delta t | T^{*} \geq t)}{\delta t} \\ &= \lambda_0(t) exp(\gamma^{t} \omega_i),\ t > 0\end{aligned} \tag{4.1}$$

where ω_i are covariates that are associated with hazard, γ is the corresponding vector of regression coefficients and $h_0(t)$ is the baseline hazard.

It is assumed that the hazard ratio $\psi = \frac{\lambda_i(t)}{\lambda_0(t)}$ depends only on covariates, whose value is fixed during the follow-up, such as gender, age, districts, previous opportunistic infection and health care facilities remain constant in the time interval between visits. However, when the interest is also in investigating whether time-varying covariates are associated with the risk for an event, the extended Cox model may be the best model to use [171]. The excellent part of this model is that it postulate that the hazard for an event, at any time point t, is associated with the extrapolated value of the covariate at the same point [172]. However, the application of time-dependent covariates is much more complicated in practice than fixed model, hence their inclusion in a survival model complicates the analysis. Furthermore, the extended Cox model is only theoretically valid for exogenous time-varying covariates because it is not appropriate when it comes to study biomarkers like viral load or cell CD4 counts. The extended Cox model is inadequate in this regard because it assumes that from one visit to the next, the biomarkers level remains the same. Hence, when the researcher fit the extended Cox model ignoring this special characteristic result in bias for the estimated effects of a biomarker [173].

Linear Mixed Effects Sub-model

The evolution of each subject in time can be described effectively by a linear mixed effect models. Each subject in the population has his own subject-specific mean response profile over time. We now re-define these models, by letting $y_i(t)$ denote the repeated measurements for the *i-th* subject (i=1,2,...,n) at time t. The measurements could be obtained at the specific time points $t_{ij}, j = 1, 2, ..., n_i$. The general linear mixed model has the form:

$$y_i(t) = \begin{cases} x_i(t)^t\beta + z_i(t)^t b_i + \varepsilon_i(t) \\ b_i \sim N(0, D), \\ \varepsilon_i \sim N(0, \sigma^2 I_{n_i}), \end{cases} \tag{4.2}$$

where X_i and Z_i are known design matrices, for fixed-effects regression coefficients β are assumed to be normally distributed with mean zero and variance-covariance matrix D, and are assumed to be independent of error terms ε_i, and σ^2 is the variance of the error terms. The linear mixed-effects models have unique feature in statistical models because they are able to account for the correlation within the measurements obtained from the same patients and it can also handle unequally spaced visits times [172]. The major challenge for the analysis of longitudinal data is the problem of missing data. Although longitudinal studies are designed to collect data on every subject in the sample at a set of pre-specified follow-up times, in practice, some subjects miss some of their planned visit for various reasons. Missing data poses several challenges in the design of longitudinal evolution studies and the analysis of data from these studies. The first statistical challenge is loss of efficiency, in the sense that the average longitudinal evolution are less precisely estimated [172, 188, 189]. Now, to compensate for that we need to enroll more patients to increase the power of the statistical test. Secondly, the missing data has the consequences of precision reduction, and as the result it affect method choices. When missing data are not properly handled, it will introduce bias and lead to misleading inferences. The missing data lead also to incomplete longitudinal responses. There exist different methods of analysis of incomplete longitudinal data and its appropriateness of different methods of analysis is determined by missing data mechanism [172]. Little *et al.*, (2002) define three types of missing data mechanism as follows:

- Missing Completely at Random (MCAR): Which postulate that the probability that the responses are missing are unrelated to observed longitudinal outcome. As a result, under MCAR we can obtain valid inference using valid statistical procedures for the data at hand, while ignoring the process generating the data. For example if a patient moves to another health care facility or forgets an appointment.

- Missing at Random (MAR): When assumed that the probability of missingness depends on set of observed responses, but unrelated to the outcome that should have been obtained. For example, the patient leaves the study on doctors advice based on previously observed longitudinal measurements.

- Missing Not at Random (MNAR): When the probability that the longitudinal responses are missing depends on observed and unobserved data. For example, a patient leaves the study due to an event, and the event is related with his HIV, including those that would have observed if they would kept attending the appointment.

4.5 Joint Model sub-structure

Joint models of longitudinal data and/or survival data have enjoyed great attention in literature over the past three decades. The importance of these models is well recognised, partly due to the fact that longitudinal data arise frequently in practice. Despite the extensive literature on joint models, these models continue to be a very active area of current Biostatistics research since they offer many advantages over separate analysis of longitudinal data and or survival data [172]. Few issues may stand out for joint models. For example, the common assumption of the distributions for the models errors and random effects in joint models is normal in most of the studies, but this assumption lacked robustness against departure from normality.

To analyse the Limpopo HIV/AIDS dataset, we will utilise the framework of joint models for longitudinal and survival data. The main purpose behind these models is to join survival model for the continuous time-to-event process with mixed-effects model for longitudinal outcome. The basic joint model is written as follows:

$$\begin{aligned} y_i(t) &= x_i^t(t)\beta + z_i^t(t)b_i + \varepsilon_i(t) \\ h_i(t) &= h_0(t)\exp[\gamma^t w_i + \alpha\{x_i^t(t)\beta + z_i^t(t)b_i\}], t > 0 \end{aligned} \quad (4.3)$$

where α quantifies the strength of the association between the marker and the risk for an event, and w_i is the baseline covariates. It is assume that the risk for an outcome dependent dropout is associated with true and unobserved value of the longitudinal outcome [172]. However, the key assumption of joint model is that the random effects underlie both longitudinal and survival process. That means these random effects account for both the association between the longitudinal and event outcome, and the correlation between the repeated measurements in the longitudinal process [173].

We postulate that joint models belong to the class of shared parameter models and define as follows:

$$P(Y_i^0, Y_i^m, T^*) = \quad P(Y_i^0, Y_i^m)P(T^*|b_i)P(b_i)db_i \quad (4.4)$$

where T^* is the true time-to-event, Y_i^0 is the longitudinal measurements before T^*, and Y_i^m is the longitudinal measurements after T^*. Thus, the association between the longitudinal process is explained by shared random effects b_i.

Stratified Joint models

It it unrealistic to always assume that the sample at hand comes from a homogeneous population. For example, HIV patients in health care facilities of Limpopo Province are different, those that receive health care in clinics are different than those receiving health care in regional hospitals. These patients are classified based on seriousness of their AIDS clinical stages. Patients in these health care facilities are assumed to be divided into different strata, with each strata having its own baseline hazard function, but common values for regression coefficients γ and a. Under the stratified relative risk joint model, the risk for patient i to belong to stratum $k$is given by:

$$\lambda_{ik}(t) = \lambda_{0k}(t)\exp\{\gamma^t w_i + am_i(t)\}$$

with $\lambda_{0k}(t)$ denoting the baseline hazard function of stratum k (k=1,...,r). The formulation of these stratified joint models assume that the effect of every covariates is constant across the strata. But, this is not always a reasonable assumption, because in many cases some covariates may have a different effect per strata. For example, it is reasonable to assume that the effect of age is the same across all health care facilities, but it will be unreasonable to assume that treatment effect is uniform across all health care facilities and also it may be less defendable [172].

Estimation of the Longitudinal Outcome

The mean, mode and median of $w_i(u|t)$ is derived in the similar manner as that of conditional survival probabilities. The mean and the mode of posterior distribution of random effects were found to be very close to each other, and therefore, we expect negligible difference between their estimators [172]. For practical purpose, the mode will be taken as preferred candidate since the posterior of our distribution is skewed. Again, obtaining standard errors of these estimators was found difficult because both random effects of the mean and mode were non-linear as a result can not be written in a closed-form, hence, Monte Carlo approach was used to derive the subject-specific mean and the median.

$$\begin{aligned} Pr(T_i \geq u|T_i > t, Y_i(t); \theta) &= \quad Pr(T_i \geq u|T_i > t, Y_i(t), b_i; \theta)p(b_i|T_i > t, Y_i(t); \theta)db_i \\ &= \quad Pr(T_i \geq u|T > t, b_i; \theta)p(b_i|T_i > t, Y_i(t); \theta)db_i \\ &= \quad \frac{S_i\{u|M_i(u, b_i, \theta); \theta\}}{S_i\{t|M_i(u, b_i, \theta); \theta\}} p(b_i|T_i > t, Y_i(t); \theta)db_i \end{aligned} \tag{4.5}$$

where S_i denotes survival function. Based on (4.16) we can derive a first-order estimate of $\pi_i(u|t)$ using the empirical Bayes estimate for b_i as follows:

$$\hat{\pi}(u|t) = S_i\{u|M_i(u, \hat{b}_i^{(t)}, \hat{\theta}), \hat{\theta}\}/S_i\{t|M_i(u, \hat{b}_i^{(t)}, \hat{\theta}), \hat{\theta}\} + O([n_i(t)]^{-1}) \tag{4.6}$$

t

where $\hat{\theta}$ denotes the maximum likelihood estimates, $\hat{b}_i$ denotes the mode of the conditional distribution $log(b_i|T_i > t, Y_i(t); \hat{\theta})$, and $n_i(t)$ denotes the number of longitudinal responses for subject i by time t. Simulation studies have shown that this estimator work very well in practice, however, deriving its standard error and confidence intervals for $\pi_i(u|t)$ is difficult due to the fact that we need to account for the variability of both the maximum likelihood and empirical Bayes estimates [184]. To resolve this problem Proust-Lima *at el.*, (2014) and Liang1 *at el.*, (1986) proposed Monte Carlos scheme as an alternative. The posterior expectation of (4.6) can be derived as follows:

$$Pr(T_i \geq u|T_i > t, Y_i(t), D_n) = \quad Pr(T_i \geq u|T_i > t, Y_i; \theta)p(\theta|D_n)d\theta \tag{4.7}$$

when sample size is large enough then $\theta|D_n$ can be approximated by $N(\hat{\theta}, var(\hat{\theta}))$. We can use Monte Carlo simulation scheme in order to calculate the first-order estimator (4.15), but using θ^{L} and b_i^{L} instead of $\hat{\theta}$ and $\hat{b}_i$,in order to propagate the uncertainty in the maximum likelihood and empirical Bayes estimates, respectively. Hence the median point estimate of $\pi(u|t)$ is given:

$$\hat{\pi}(u|t)) = median\{\pi_i^{l}(u|t), l = 1, ..., L\} \tag{4.8}$$

and average point estimate is given by

$$\bar{\pi}(u|t)) = L^{-1} \sum_{l=1}^{L} \pi^{l}(u|t) \tag{4.9}$$

4.5.1 Joint Latent Class Models

There is another class of joint models called joint model latent class (JLCM) [180, 187]. The motivation behind this type of joint models is to account for possible heterogeneity in the population, and it assumes that the sub-populations that constitute population are latent, in the sense that heterogeneity is not captured by any of the observed covariates [172]. This assumption of heterogeneity is quite relevant in medical research where different profiles of patients are expected. For example, how AIDS progress after treatment, we normally observe different profiles of HIV+ patients. Although little attention has been given to JLCM in research, yet, JLCM offers computationally attractive alternative to the shared random effect model (SERM) and is based on different assumptions that link between the longitudinal and time-to-event component of the model [180]. The main difference between SERM and JLCM is that in SERM the link must be precisely defined than in JLCM. In our study we shall not describe other special joint models such as joint models based on pattern-mixture modelling or simple transformation models.

Latent class membership probability

In the latent class joint models, we assume a population of N subjects that can be divided into a finite number G of latent homogeneous sub-populations, and furthermore, there is unobserved class indicator $c_i = 1, ..., G$ (a categorical latent variable) that denotes the class membership of the *i-th* subject, which equals g if subject i belong to latent class g (g =1,...,G). An individual has a probability π_{ig} of belong to latent class g, which is modelled using a multinomial logistic regression according to covariates X_π [180]:

$$\pi_{ig} = \frac{\exp^{\xi_{0g}} + X_\pi^t \xi_{lg}}{{}_{l=1}^{G} \exp^{(\xi_{0l} + X_\pi^t \xi_{1l})}} \tag{4.10}$$

where ξ_{0g} is the intercept for class g and ξ_{1g} is the vector of class-specific parameters associated with the vector of time-independent covariates X_π . For identifiability, $\xi_{0l} = 0$ and $\xi_{1l} = 0$

Class-specific marker trajectory

Each latent class is characterised by a class-specific marker trajectory. Now, given the latent class g, the vector of repeated measures of the longitudinal marker $Y_i = (Y_i(t_{ij}), ..., Y_i(t_{in_i}))$ is described at different times of measurement t_{ij} $(j = 1, ..., n_i)$ by a standard linear mixed model [134]:

$$Y_i(t_{ij}) | c_i = g = Z_i(t_{ij})^t u_{ig} + X_{li}(t_{ij})^t \beta_g + \varepsilon_i(t_{ij}) \tag{4.11}$$

where the p-vector of class-specific random-effects $u_{ig} = u_i | c_i = g \sim N(\mu_g, B_g)$. The n_i-vector of measurement error $\varepsilon_i \sim N(0, \Sigma_i)$. The variance-covariance matrix B_g can be common over classes or class-specific. However, when you consider class-specific, usually $B_g = \omega^2 B$ with B unstructured and $\omega_G = 1$ to limit the number of parameters and identifiability concerns. The variance-covariance matrix Σ_i is usually restricted to $\sigma^2 I_{n_i}$for homoscedastic independent errors, however, ε_i can also be include a correlation process such as a Brownian motion or auto-regressive process. No overlap between $Z_i(t_{ij})$ and $X_{li}(t_{ij})$ is assumed for identifiability [180].

Class-specific risk of event

We let T^* denote the time-to-event of interest, C_i the censoring time, $T = min(T^*, C_i)$ and $E_i = 1$ for $T^* \leq C_i$. Now given the latent class g, the risk of an event described using any survival model, or for simplicity we can consider a proportional hazard model [180]:

$$\lambda_i(t|c_i) = g(\zeta_g, \delta_g) = \lambda_{0g}(t, \zeta_g) \exp^{X_{ei}(t)^t \delta_g} \tag{4.12}$$

where $X_{ei}(t)$ is the r-vector of covariates associated with the r-vector of parameter δ_g . The class-specific hazard stratified on the latent class structure or baseline hazard proportional in each latent class can be considered. The only parametric hazard functions such as Weibull, piecewise-constant or M-splines [180].

Posterior Classification

The posterior classification can be obtained from the posterior estimates of the latent class membership probabilities as follows:

$$\begin{aligned} p(T_i, \delta_i, y_i|c_i = g, b_i; \theta) &= p(T_i, \delta_i|c_i = g; \theta)p(y_i|c_i = g, b_i; \theta) \\ p(y_i|c_i = g, b_i; \theta) &= \prod_j p(y_i(t_{ij})|c_i = g, b_i; \theta) \end{aligned}$$

These models assume that the correlations between the repeated measurements in the longitudinal outcome are captured by the random effects b_i, whereas the association between the event time and longitudinal processes is explained by the shared latent class indicator c_i. The major advantage of this type of model is that it allows for more flexible association structure compared to the classical joint models that assumes the same set of random effects b_i to account for both types of association [184]. Under the above conditional independence assumptions, a general latent class joint model is defined as follows:

$$\begin{cases} h_i(t|c_i) = h_{0g}(t)\exp(\gamma_g^t w_i), \\ y_i(t|c_i = g) = x_i^t(t)\beta_g + z_i^t(t)b_{ig} + \varepsilon_i(t), \\ Pr(c_i = g) = \exp(\lambda_g^t u_i) / \sum_{l=i}^{G} \exp(\lambda_g^t u_i), \end{cases}$$

where u_i denote the vector of covariates associated with these probabilities with corresponding vector of regression coefficients vector $\lambda^t = (\lambda_i^t, \ldots, \lambda_G^t)$, with $\lambda_G = 0$ for identifiability. The random effects $b_{ig} \sim N(\mu_g, \sigma_g^2 D)$ are assumed to be latent-class specific, and their covariance matrix is assumed to depend on c_i only via the scalar variance parameter σ_g^2.

It is postulated that postulate that patients in different latent groups have both different longitudinal evolutions and different risks for an event [186].

Using the fitted model we can derive the posterior classification for the patients in the sample, and to achieve that we take the maximum of the posterior probabilities as follows:

$$Pr(c_i = g|T_i, \delta_i, y_i; \hat{\theta}) = \frac{Pr(c_i = g; \hat{\theta})h_i(T_i|c_i = g; \hat{\theta})^{\delta_i} S_i(T_i|c_i = g; \hat{\theta})p(y_i|c_i = g; \hat{\theta})}{\sum_{l=1}^{G} Pr(c_i = l; \hat{\theta})h_i(T_i|c_i = l; \hat{\theta})^{\delta_i} S_i(T_i|c_i = l; \hat{\theta})p(y_i|c_i = l; \hat{\theta})}$$

that is subject i is classified to group g, using $\hat{c}_i = arg\max[Pr(c_i = g|T_i, \delta_i, y_i; \hat{\theta})]$

which is similar in spirit to empirical Bayes estimates for the random effects [187].

4.5.2 Accelerated Failure Time Joint Models

The accelerated failure time (AFT) joint models specify that the predictors act multiplicatively on the failure time or additively on the log failure time. The accelerated failure time joint model is defined as follows:

$$log(T_i) = \gamma^t w_i + \sigma_t \varepsilon_{ti} \tag{4.13}$$

where σ_t is a scale parameter and ε_{ti} is assumed to follow certain distribution, γ^t denotes the change in the expected log failure time for a unit change in the corresponding covariate w_{ij}. Equivalently, a unit change in ω_{ij} increases the failure time by a factor of $\exp(\gamma_j)$. Weibull distribution T_i is the only parametric distribution that accepts both a relative risk and an AFT model formulation.

Dynamic Predictions for Survival Probabilities

During a follow-up, for a specific patient and at a specific point, we would like to utilise available information we have to predict survival probabilities. This information is vital to physicians to gain better understanding of the disease dynamics, and ultimately take optimal decision at that specific time point. In these technological era, there is a trend in medical practice towards personalised medicine, and there is a prominent role such individualised predictions can play in that respect, as a result there has been a lot of interest within joint modeling framework in that front [180, 182, 179, 181].

Rizopoulos (2012) put it more formally, based on a joint model fitted in a random sample $D_n = \{T_i, \delta_i, y_i; i = 1, ..., n\}$, wherein we are interested in predicting survival probabilities for a new subject i that has provided a set of longitudinal measurements $Y_i = \{y_i(s); 0 \leq s < t\}$ and has a vector of baseline covariates w_i.

Since $y_i(t)$ has an important characteristic of the endogenous nature, it means that it is directly related to the failure mechanism. Hence, it is more relevant to focus on the conditional probabilities of survival time $u > t$, given survival up to t. That is,

$$\pi_i(u|t) = Pr(T_i^* \geq u|T_i > t, Y_i(t), w_i, D_n; \theta^*), \quad t > 0, \tag{4.14}$$

where θ^* denotes the true parameter values. Clearly, from (4.14), when a new information is recorded for patient at time $t^{\dagger} > t$, we can update these predictions and obtain $\pi_i(u|t^{\dagger})$, with $u > t^{\dagger}$, and therefore proceed in a time dynamic manner.

Dynamic Predictions for Longitudinal Outcomes

Very often interest may lie in the prediction for longitudinal outcome. For example, in HIV-infected patients the CD4 counts and viral load are often used to determine when treatment should be initiated. In this setting it is critically important and helpful to the treating physician to gain an insight into projected individual profile of the markers in order to initiate treatment sooner than later.

For specific subject i who is still alive by follow-up time, Rizopoulos (2012) proposed that the expected value of longitudinal outcome at time $u > t$ given his time point $Y_i(t) = \{y_i(s); 0 \leq s < t\}$ is given by:

$w_i(u|t) = E(Y_i(u)|T_i > t, y_i(t), D_n, \theta\}$, $u > t$. Similarly to the conditional probabilities (4.14), these predictions are dynamically updated in time as new additional information is recorded for that subject.

4.6 Data Analysis

Table 4.1: Parameter Estimate for Extended Cox Model with time dependent covariates

Parameter	coef	exp(coef)	se(coef)	z-value	P-value
GENDERMale	0.412694	1.510882	0.039784	10.373	< 0.0001
PrevOI	0.2229941	1.258525	0.061708	3.726	0.000194
Mopani	0.079145	1.082361	0.050848	0.1.557	0.1119586
Sekhukhune	0.036738	1.037421	0.060670	0.606	0.544825
Vhembe	0.054984	1.056523	0.0666457	0.827	0.408036
Waterberg	0.039727	1.040527	0.078610	0.505	0.613298
CD4	-0.030048	0.970399	0.0033385	-8.877	< 0.0001
AGE	0.018461	1.018632	0.001365	13.527	< 0.0001
ClinicalStage	0.254948	1.290395	0.021227	12.011	< 0.0001
Clinics	-0.482472	0.617256	0.078295	-6.162	< 0.0001
District Hospitals	-0.072037	0.930496	0.0616696	-1.168	0.242967
Regional Hospitals	-0.243628	0783779	0.131121	-1.858	0.063163
Provincial Hospitals	-0.594715	0.551720	0.200768	-2.962	0.003055
Psychiatric Hospitals	-0.027432	0.972941	0.192086	-0.143	0.886439
$VLlog10$	0.289777	1.336129	0.016631	17.424	< 0.0001

Table 4.2: Parameter Estimate for Joint Model

		Standard	95%	
Parameter	Value	Error	Confidence Interval	P-value
GENDERMale	0.3157	0.0408	(0.2356;0.3957)	<0.0001
PrevOI	0.2092	0.0627	(0.08643;0.3320)	0.0008
Mopani	0.0457	0.0515	(-0.0552;0.1465)	0.3747
Sekhukhune	0.0627	0.0614	(-0.0577;0.1831)	0.3076
Vhembe	0.0543	0.0673	(-0.0776;0.1862)	0.4193
Waterberg	0.0437	0.0795	(-0.1120;0.1995)	0.5820
CD4	-0.0247	0.0034	(-0.03134;-0.0180)	<0.0001
AGE	0.0227	0.0015	(0.0197;0.0258)	<0.0001
ClinicalStage	0.2111	0.0216	(0.1688;0.2534)	<0.0001
Community Health Centre	-0.4276	0.0790	(-0.5824;-0.2728)	<0.0001
District Hospitals	-0.0875	0.0626	(-0.2102;0.0352)	0.1626
Regional Hospitals	-0.2374	0.1322	(-0.4965;0.0217)	0.725
Provincial Hospitals	-0.5743	0.2031	(-0.9724;-.1763)	0.0047
Psychiatric Hospitals	0.0364	0.1948	(-0.3454;0.4183)	0.8517
Association Parameter (a)	0.5067	0.0599	(0.3893;0.6242)	<0.0001

We start our analysis by fitting the extended Cox model in which the logarithmic viral load is taken as exogenous time-dependent covariate. In the model

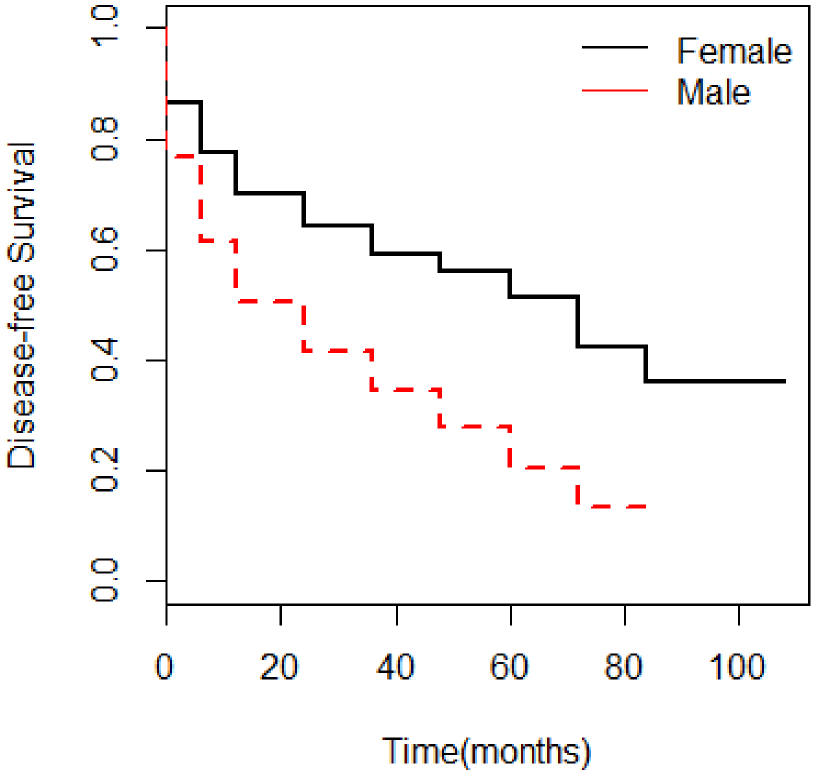


Figure 4.1: Kaplan-Meier estimator of event-free survival probabilities for for females and males.

terms we have $\lambda_i(t) = \lambda_0(t)exp[\gamma^t w_i + ay_i(t)]$ where $y_i(t)$ denotes the observed level of the logarithmic viral load and w_i are fixed covariates. Parameter a quantifies the association between features of the marker. Contrary to the Cox model where $\lambda_0(t)$ is unspecified, here the baseline risk function is assumed piecewise-constant with three knots placed at equally spaced percentiles of the observed event times. In Table 4.1, the survival sub-model, we observed that viral load has indeed a strong association with the risk for death. A unit increase in the logarithmic viral load corresponds to a $exp(-a) = 0.74 - fold$ increase in the risk for death (95% CI: 1.2933; 1.3804). Figure 4.1 depicts the Kaplan-Meier estimates of event-free survival for females and male groups.

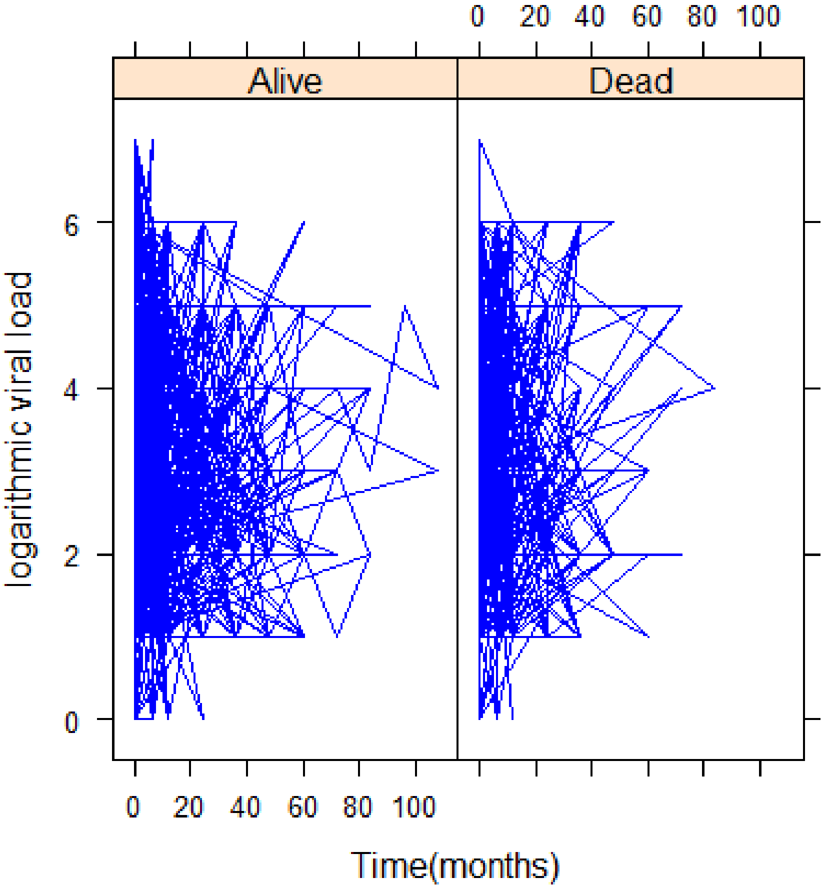


Figure 4.2: Subject-specific longitudinal trajectory for logarithmic viral load for patients with and without an event .

Clearly, females enjoy much event-free survival as compared with males.

In order to fit our joint model using JM package in R, we need first to fit separately the linear mixed-effects and Cox models, and supply the returned objects as main arguments in the function *jointModel*(). We proceed by specifying and fitting a joint model that explicitly accounts for endogeneity of the viral load marker. We first fit the linear mixed-effects model $y_i(t) = \beta_0 + \beta_1 t + \beta_i t^2 + b_{i0} + b_{i1} t + \varepsilon_i(t)$ where in the fixed effects part, main effects were icluded, and in the random-effects design matrix we included an intercept and time term. In Figure 4.2, we investigated viral load and we observed that for any individual there seem to be linear. Hence in the mixed-effects model we would allow flexibility in the specification of these profiles in both fixed-effects and random-effects.

The joint model in Table 4.2, we found a strong association between the viral load and the risk for death, with a unit increase in $log(Viralload)$ corresponding to a $exp(-a) = 0.6 - fold$ in the risk for death (95% :0.38927;0.62420). We will use Wald test for testing whether each of the fixed effects β in the longitudinal sub-model are statistically different from zero, that is,

Table 4.3: A comparison of joint and Cox extended models

	Extended Cox Model	Joint Model
variable	log Hazard Ratio(se)	Log Hazard Ratio(se)
Viral load	0.2811(0.01531)	0.0159(0.0354)
Assoct		0.3897(0.0529)

$$H_0 : \beta_1 = \beta_2 = ...\beta_6 = 0 \qquad (4.15)$$

$$H_A : \beta_j \; /= 0, j = 1, .., 6.$$

The results in Table 4.4 indicate a strong overall time effects. However, the problem with Wald test for testing fixed effects of longitudinal sub-model is that it is based on standard errors which underestimate the true variability in $\hat{\beta}$ because they do not take into account the variability by estimating the

Table 4.4: Wald test for longitudinal process

Variable	Chisq	df	Pr(> $\|Chi\|$)
Time	64.1409	4	< 0.0001
GENDER	7.3514	1	0.0067
PrevOI	0.5217	1	0.4701
Age	49.5267	1	< 0.0001
Time:GENDER	7.3514	1	0.0067
TIME:PrevOI	0.5217	1	0.4701
Time:Age	49.5267	1	< 0.0001

variance components, that is, the variance matrix for random effects [185]. In joint models this problem could be exaggerated because we do not only ignore the fact that we estimate the variance components, but also that we need to estimate survival process. Hence, likelihood ratio tests are preferred in joint modelling.

In order to implement likelihood ratio test in joint models, we need first fit the joint model under null hypothesis, that is, joint model with no covariates effect in the survival sub-model. Table 4.5 shows that indeed there is association between logarithmic viral load and risk for death for the advanced HIV infected patients of Limpopo Province AIDS dataset.

Table 4.5: Likelihood Ratio test for Joint Models

Joint model	AIC	BIC	log.Lik	LRT	df	P-Value
Joint Model	38291.61	38377.15	-19133.80			
Full Joint Model	38212.34	38305.01	-19093.17	81.27	1	< 0.0001

Table 4.6: Likelihood Ratio Test

	AIC	BIC	log.Lik	LRT	df	P-value
Basic Joint Model	38293.45	38386.12	-19133.72			
Full Joint Model	36203.33	36495.6	-18060.66	2146.12	28	< 0.0001

Table 4.6 we tested as whether the covariates in the survival sub-model contributes something in explaining the variability in the risk for death of advanced AIDS patients in Limpopo Province. Since AIC and BIC are smaller for full model, we conclude that there is an association between the logarithmic of viral load and the risk for death for advanced HIV+ infected patients of Limpopo AIDS dataset, and the covariates in the survival sub-model contributes something in explaining the variability in the risk for death of the advanced AIDS patients, that is, the covariates have significant effects contribution in the risk for death.

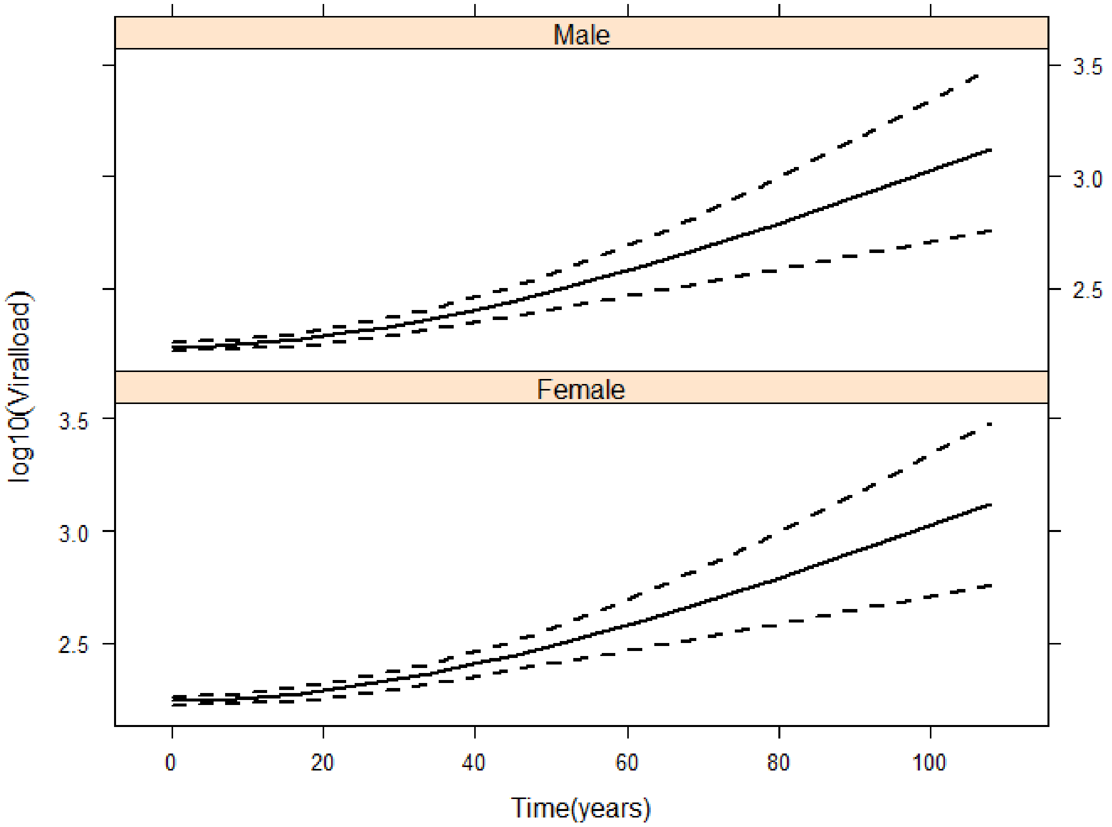


Figure 4.3: Fitted average longitudinal profiles of male and female for viral load for HIV patients with median age.

Latent Class Joint Models

Table 4.7: Posterior classification based on longitudinal and time-to-event data

	Class 1	Class 2	Class 3	Total
N	357	8231	627	9215
%	3.87	89.32	6.8	1.00

Table 4.7 demonstrates that class 2 (89.32%) contains the largest percentage of subjects , followed by class3 (6.8%) and thereafter class 1 (3.87%).

In Table 4.8, the higher means of posterior probabilities (class 1: 0.8695; class 2:0.9927 ; class 3:0.9492) for each class suggest that for the majority of patients class allocation was evident.

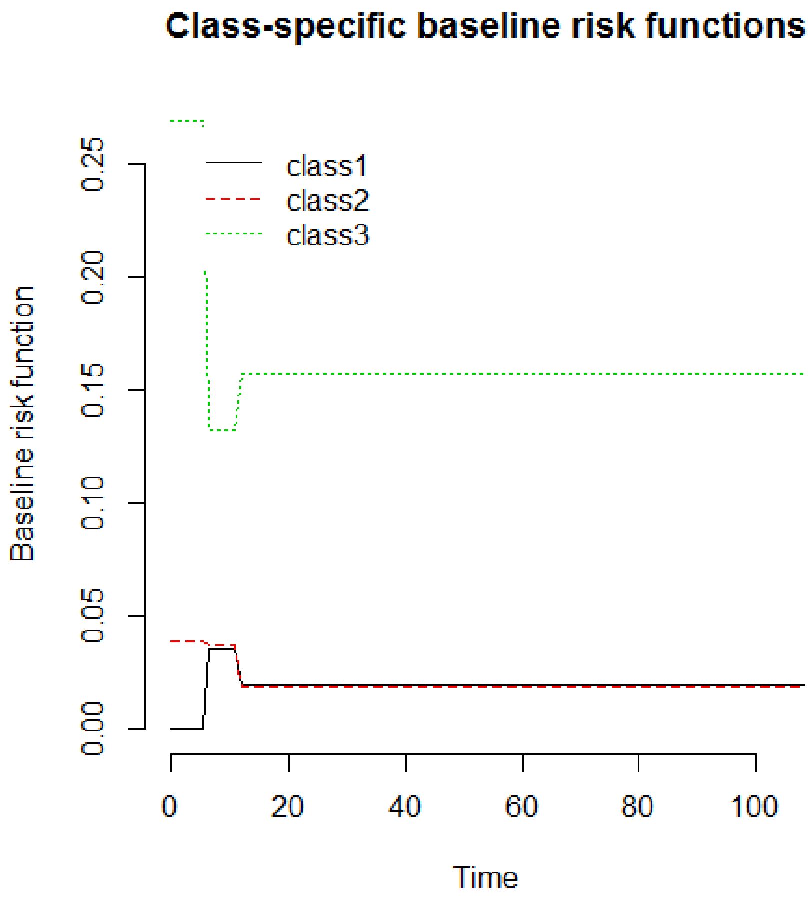


Figure 4.4: Fitted average longitudinal evolutions for three class joint model of Limpopo AIDS dataset .

Table 4.8: Mean of posterior probabilities in each class

Classes	Prob 1	Prob 2	Prob 3
Class 1	0.8695	0.0460	0.0845
Class 2	0.0072	0.9927	0.0001
Class 3	0.0467	0.0041	0.9492

Figure 4.4 illustrates the average longitudinal evolutions of the logarithmic viral load for three latent classes, and Figure 4.5 show their corresponding event-free survival probabilities. It is quite evident that the model has identified three distinct sub-populations. In particular, class 1 with relatively high

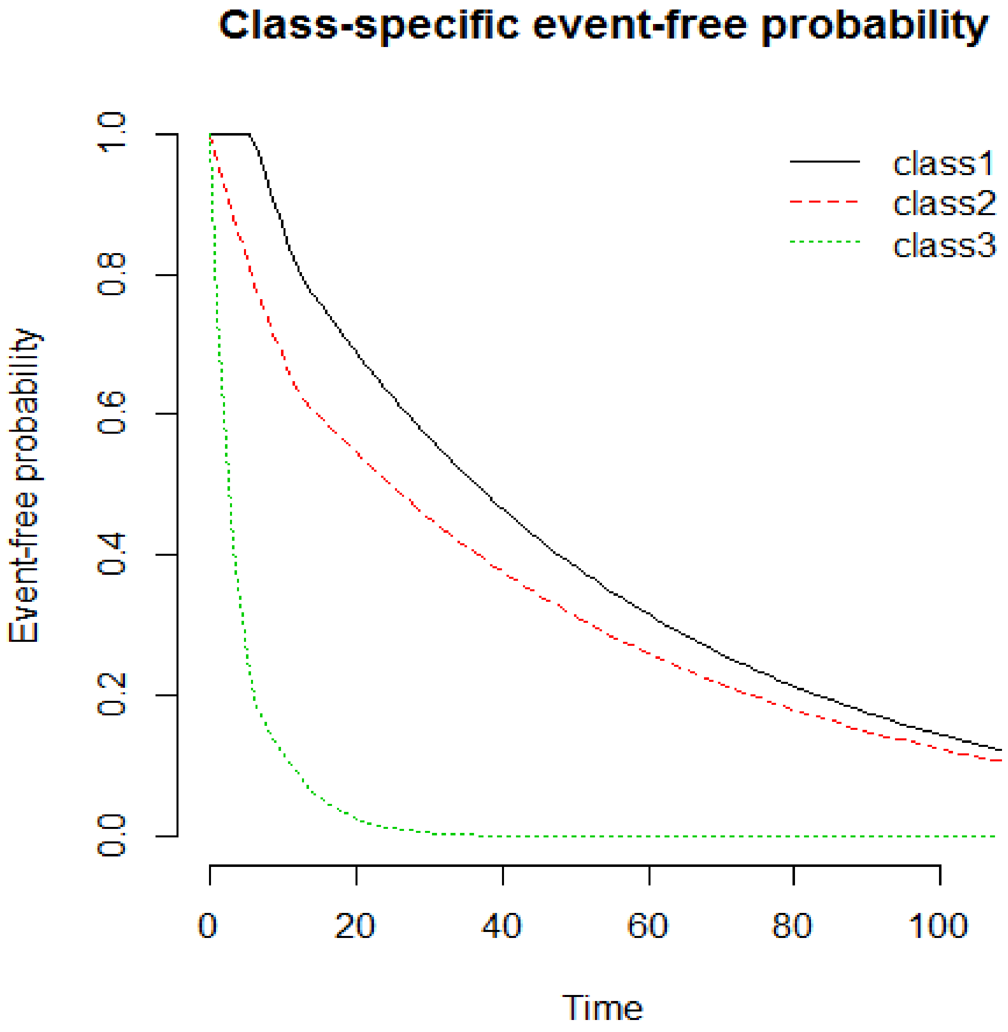


Figure 4.5: Event-free survival probabilities for three class joint model of Limpopo AIDS dataset.

viral load levels, however, it does not enjoy high event-free survival rates. That is, in contrary, it has lowest event-free survival rate. Class 2 starts with relatively high viral load levels and drop to stable viral load levels, and thereafter enjoy a relatively moderate event-free survival rates. Class 3 starts with low viral load levels and then go up for short space of time and thereafter drop for stable viral load levels corresponding to the highest event-free survivalrates.

Accelerated Failure Time Models

Table 4.9: Comparison between accelerated relative risk and accelerated failure time models

Model	AIC	BIC	log.Lik
Basic joint Model	45491.02	45562.31	-22735.51
AFT Joint Model	48252.16	48323.45	-24116.08

Since these models in Table 4.9 are not nested, they can be compared using information criteria. We observe that basic joint model has a smaller AIC and BIC, hence it is a preferred model. The parameter estimates and standard errors for the event process under the two models are presented in Table 4.10. Evidently, the estimated regression coefficients for the estimated covariates have different values and as well as opposite signs under the two joint models due to the fact that they have different interpretations.

Table 4.10: Parameter estimates and standard error for the Weibull model fitted to the AIDS dataset under the relative risk and accelerated failure time formulations

	Relative Risk		**AFT**	
Parameter	Value	Std.Error	Value	Std. Error
Intercept	-0.9837	0.0813	17.9838	0.5632
GENDERMale	0.5000	0.0388	-3.2208	0.2500
Assoct	0.3657	0.0230	-4.0355	0.1815
log(shape)	-2.9240	0.0605	-1.8299	0.0183

4.6.1 Joint Model diagnostics

Residuals for the longitudinal part

It is essential that joint model assumptions are validated, and the standard tool to assess these assumptions are residuals plots. In particular, in the longitudinal it is assumed that linear subject-specific evolutions in time for the logarithmic viral load, whereas in the survival part, the effect of true viral loads is considered, and assume a piecewise-constant baseline risk function. For the fitted joint models these residuals for the longitudinal part are illustrated in Figure 4.7, and it includes the plots of the subject-specific residuals versus the corresponding fitted values, the Q-Q plot of the subject-specific residuals, and the marginal survival and cumulative risk functions for the event process.

Residuals for the survival part

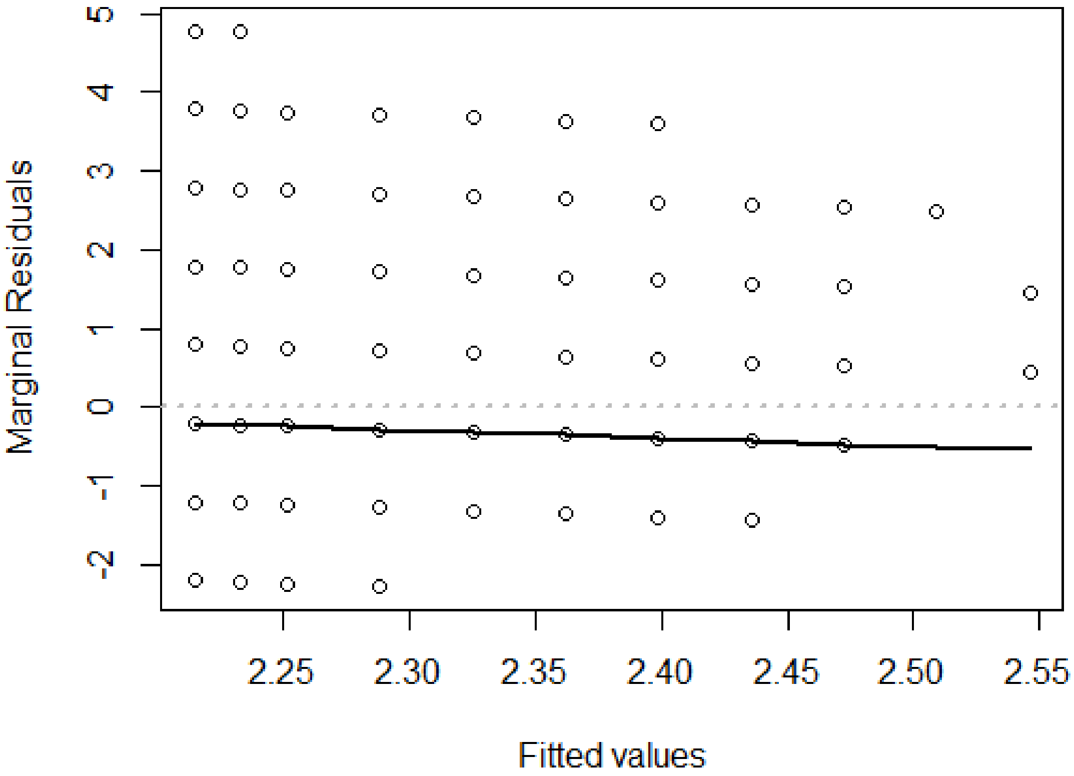


Figure 4.6: Marginal standardized residuals versus fitted values for the longitudinal outcome for the Limpopo Province AIDS dataset.

We observe in Figure 4.6 that for smaller fitted values we have more positive than negative. Small fitted values corresponds with higher levels of logarithmic viral loads , which in turn corresponds to a worsening of patient's condition and therefore higher chances of being censored. Thus, the residuals corresponding to smaller values are only based on patient's with a good health condition. Hence, due to dropout one can not differentiate with certainty that the systematic trend observed in Figure 4.6 can be attributed to a mis-specification of the design matrix X of fixed effects [184]. Figure 4.10 shows a scatter-plot with superimposed loess curve. The grey solid line denotes the fit of loess smoother. Clearly, one can observe from the fitted values that there is huge deviation of the loess smoother from zero. When plot both the plots as well as loess smoother are plotted the result is shown in Figure 4.11. Great deviation from zero is observed for both females and males. We proceeded our residuals analysis for survival outcome by assessing the overall fit on the survival sub-model using Cox-Snell residuals illustrated in Figure 4.9. The black solid line denotes Kaplan-Meier estimates of the survival functions of the Cox-Snell residuals with gender, and the grey solid line, the survival function. We observe lack of fit for the residuals values greater than0.5.

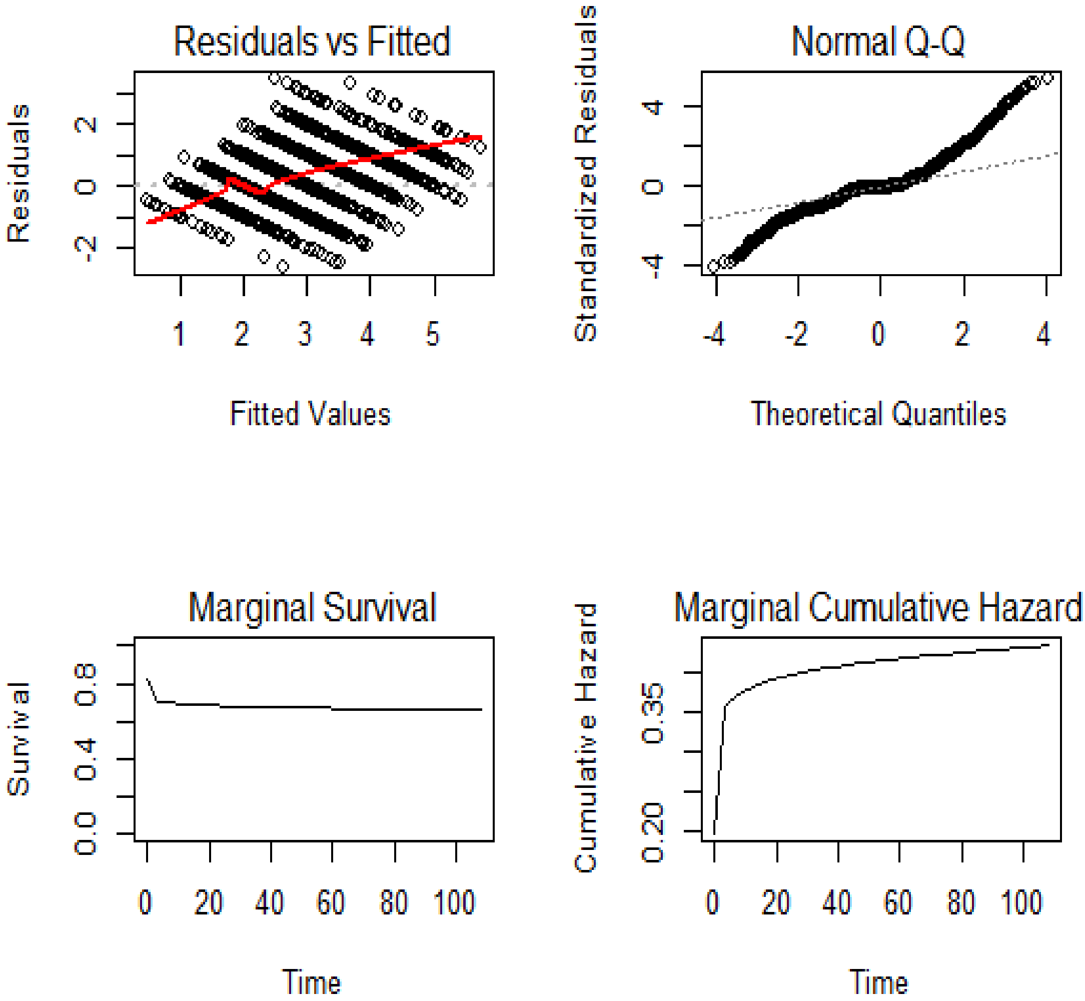


Figure 4.7: Diagnostic plots for the joint model fitted to Limpopo AIDS dataset.

4.6.2 Dynamic prediction

In equation 4.14 the patient had survived up to the last point t on which the viral load was recorded, and will produce survival probabilities for a set of predefined $u > t$ values, and the output is shown in Table 4.11. In fact Table 4.11 are regular sequence of equidistant points from the minimum to the maximum observed event time as well as computed $\pi_i(u|t)$ for $u > t$ in the sequence. The first row in our output in Table 4.11 corresponds to the last time point for which

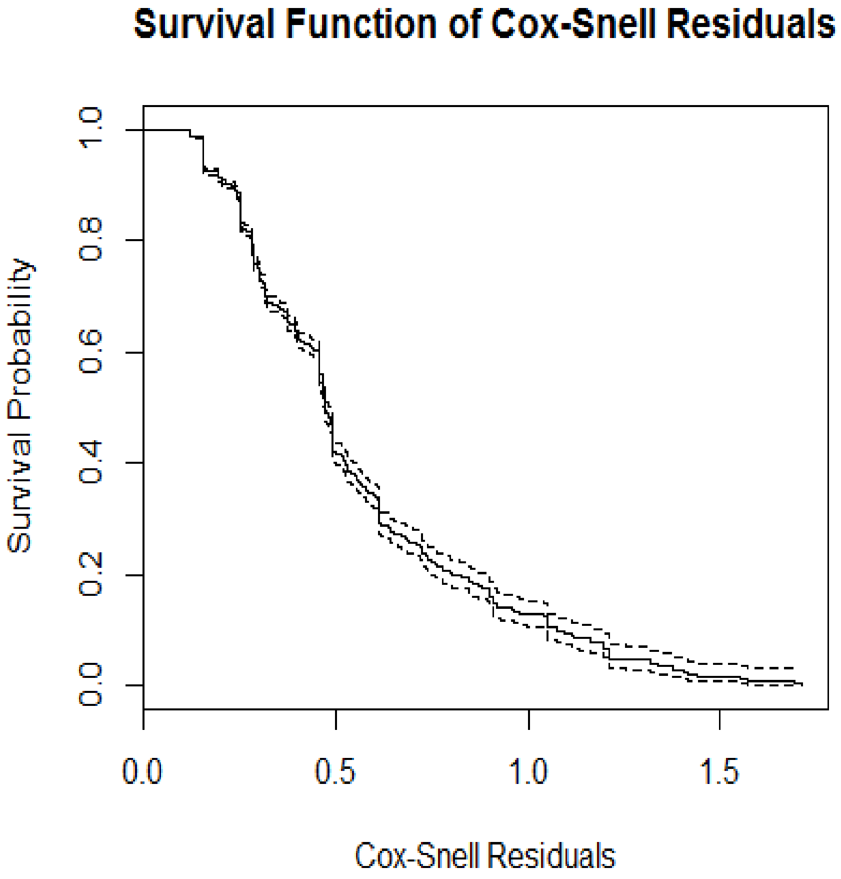


Figure 4.8: Cox-Snell residuals.

patient 18 was still event free and the corresponding estimates and 95% confidence interval.

Table 4.13 shows only the point estimates ($\pi_i(u|t)$) . When the Tables 4.11 and 4.12 are compared, negligible difference is observed. However, Table 4.12 yields a better accurate results because they properly approximate the integrals in the definition of $\pi_i(u|t)$. Tables 4.11 and 4.12 are true for any patients in our study.

Figure4.14 depicts the survival probability of patient 18 from Limpopo dataset.

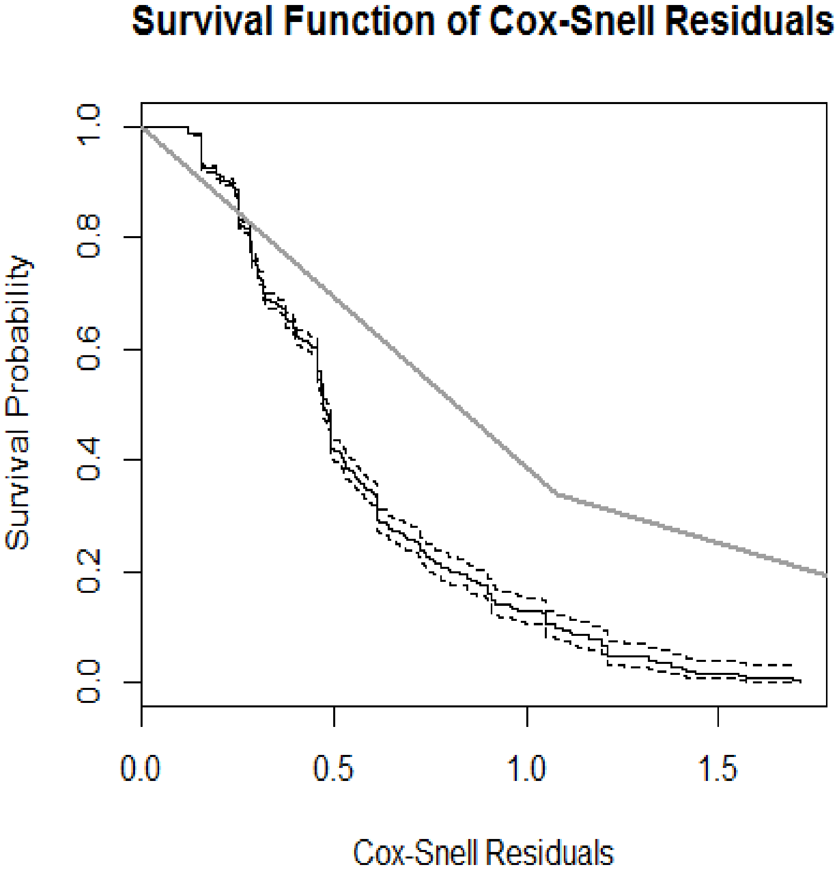


Figure 4.9: Cox-Snell residuals for Limpopo AIDS dataset.

The red solid and dashed lines correspond to the mean and median estimators, respectively, and the 95% confidence pointwise intervals. The vertical dotted line in Figure 4.13 represent the time point of the last viral load measurements. The right of the vertical line, the solid line represent the median estimator for $\pi_i(u|t)$, and dotted lines correspond to the 95% pointwise confidence intervals. We observed that after the first (t=0) measurement the viral load increased while the rate of the conditional survival probabilities were decreasing.Clearly the health condition of the patient 18 was deteriorating drastically.

We assume that in (4.14) the patient had survived up to the last point t on which the viral load was recorded. Table 4.13 shows the conditional probabil-

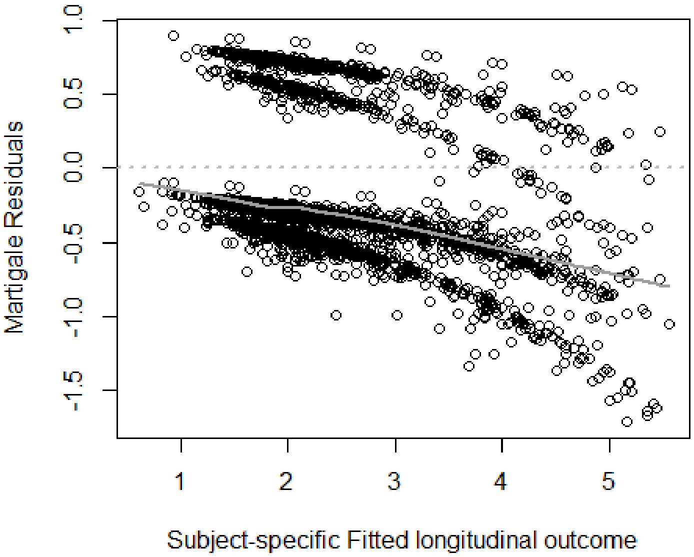


Figure 4.10: Martingale residuals versus the longitudinal outcome subject-specific fitted values of longitudinal outcome for Limpopo AIDS dataset.

ities for event of patient 18, and at 27 months she was sill event free, and it was the last time her viral load was recorded. For this patient the conditional probability that she will still be alive at month 28 is 0.9392 for the mean and 0.9431 for the median, (95%:(0.8908;0.9681)). In our study we would prefer the median conditional probability because our posterior distribution is skewed.

Dynamic Predictions for the Longitudinal Outcome

In Figure 4.15 each panel denotes the time point of the longitudinal measurements recorded for patient 18. The red points denote the predicted longitudinal trajectory, while the green points denote the 95% confidence intervals. The prediction for patient 18 are updated when additional log viral load measurements are recorded. The elaborative plot is shown in Figure 4.15. It was observed that the width of the predictions intervals become wider and wider as time progress, indicating that we have much believe that on prediction shortly after the last

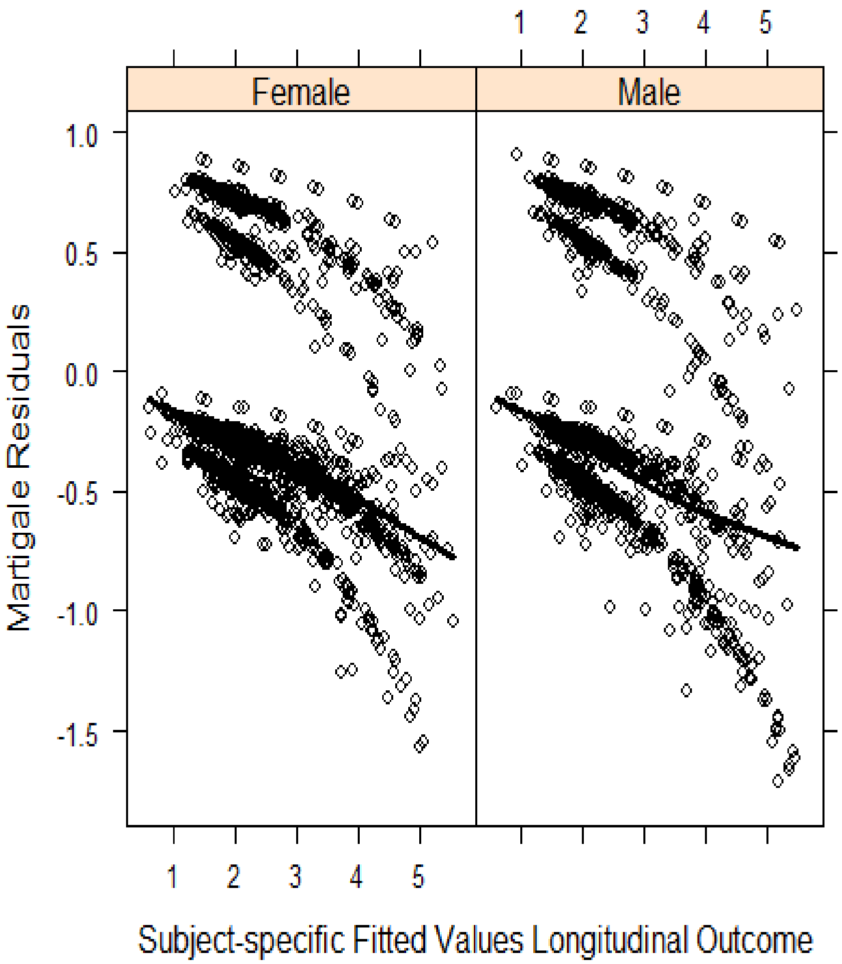


Figure 4.11: Martingale residuals versus the subject-specific fitted values per gender for Limpopo AIDS dataset.

recorded available viral loads measurements. According to Rizopoulos (2011) the important features of these predictions intervals is that they are not restricted to be symmetric, since they are not based on an asymptotic normality. We expect Monte Carlo approach to provide a relatively good approximation to the true sampling distribution, and therefore obtain confidence intervals that have higher probability to satisfy the claimed intervals [184].

Chapter 4 the following objectives were addressed: separate longitudinal and

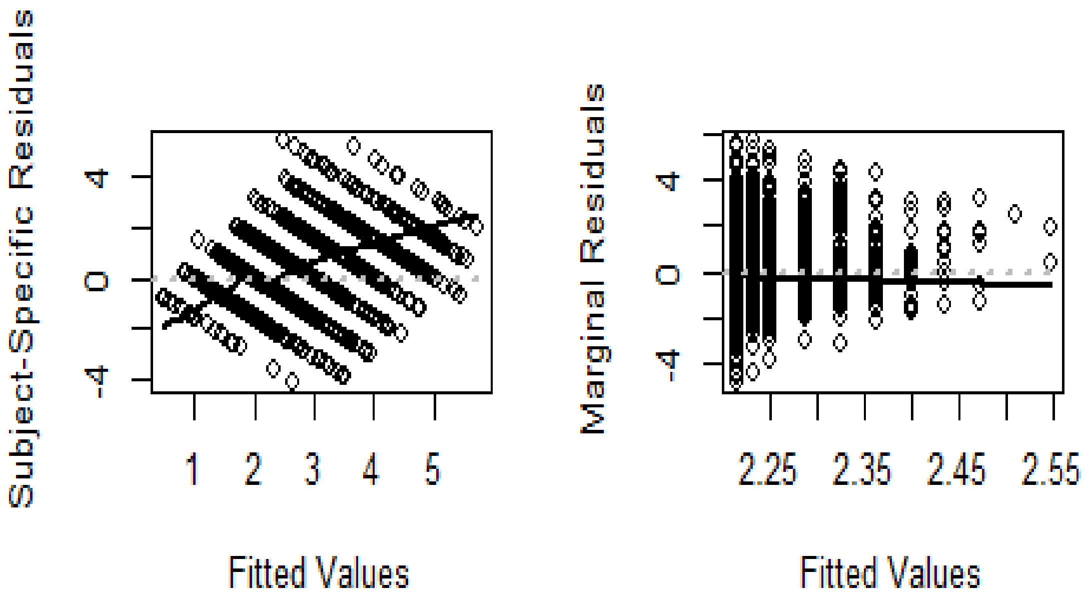


Figure 4.12: Scatterplot of observed residuals versus fitted values for longitudinal process for Limpopo AIDS dataset.

survival analyses per outcome was performed; the strength of association between the longitudinal evolution of viral load and hazard rate to death was established; separate and joint models , and various association measures such as parametric joint and shared parameter models approach were compared; Fu482theremore, the average evolutions between males and females were compared; how marker-specific evolutions are related to each other (association of the evolution) was realised; prediction for time to death for any randomly selected HIV positive patient by considering patient's viral load was computed; and came up with good joint model(s) that will handle simultaneously both the repeated measurements as well as the survival outcomes in the presence of clustering in the Limpopo Province HIV/AIDS dataset.

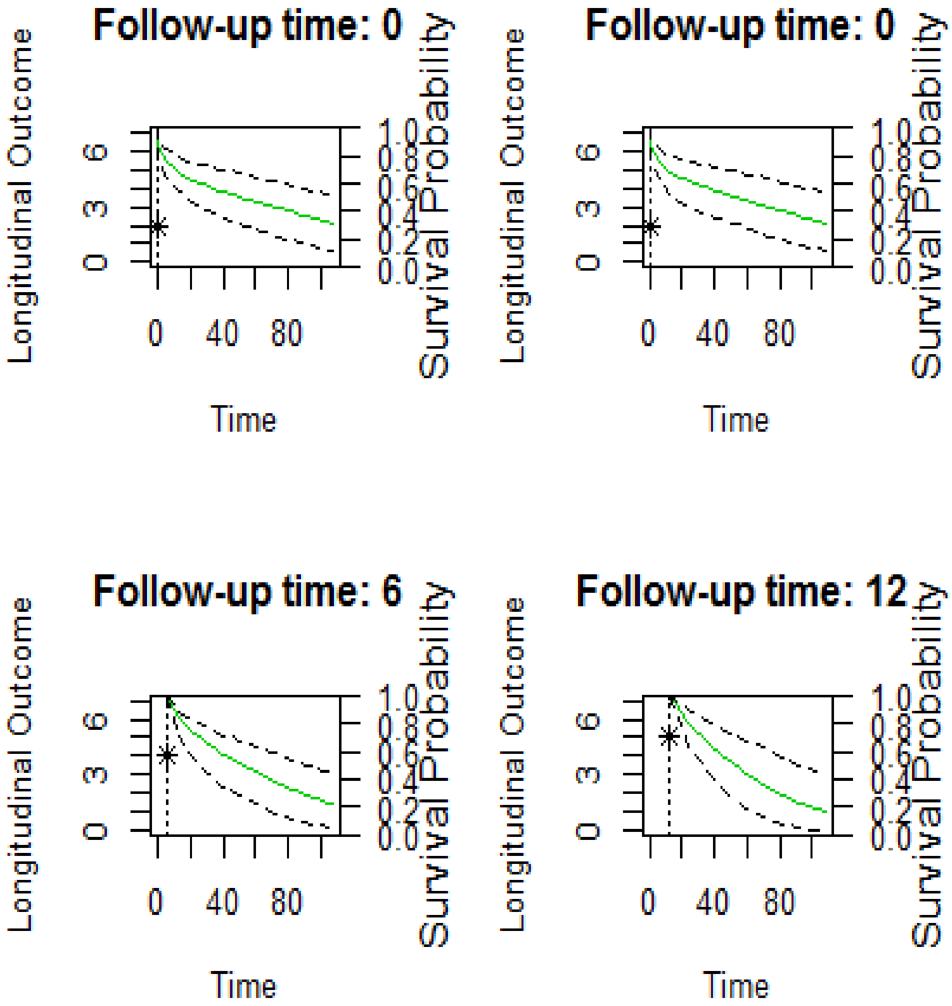


Figure 4.13: Dynamic survival probabilities for patient 18 from Limpopo AIDS dataset during follow-up.

4.7 Discussion

The researcher in HIV/AIDS study in Limpopo Province, employed joint model using secondary dataset of Department of Health showed the benefits of joint modelling when both the longitudinal and survival processes are associated with unknown covariates. The time-dependent Cox model provided a naive estimates of how individual's viral load levels affect their survival and were

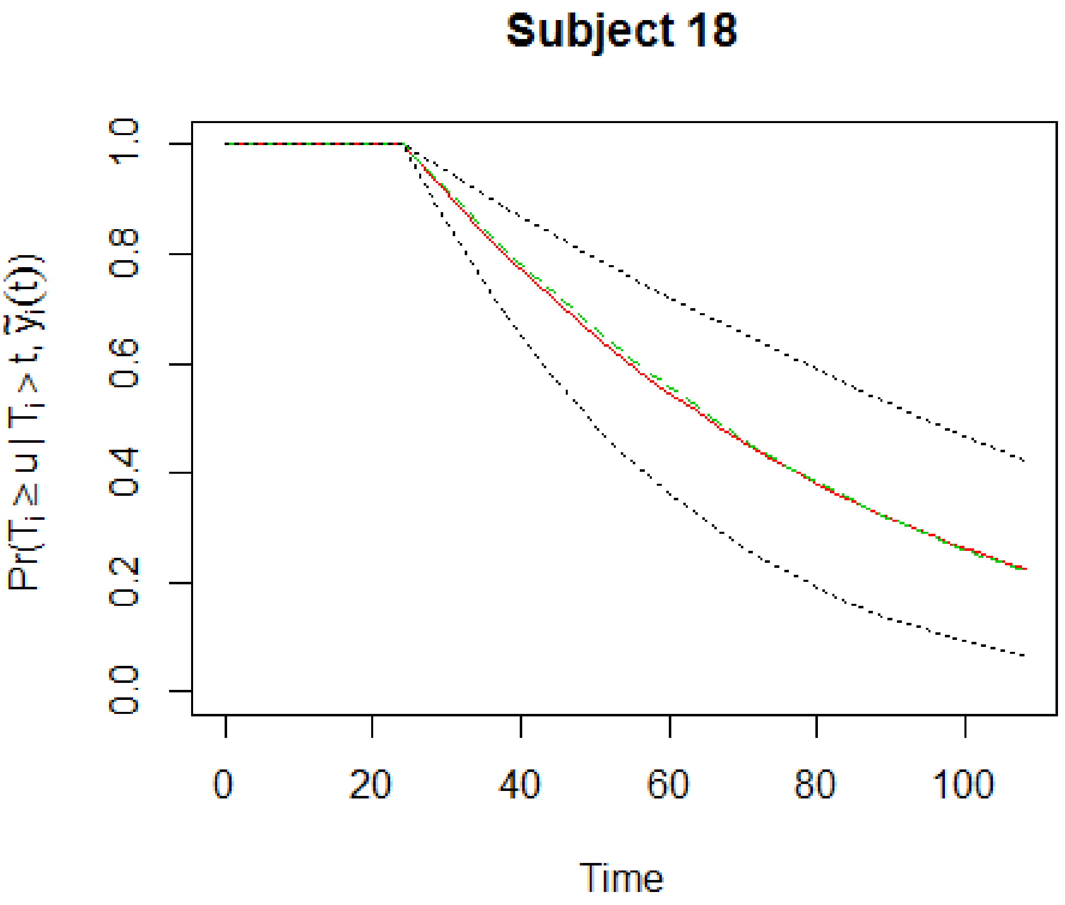


Figure 4.14: Survival probability for patient 18.

compared with joint model analysis results. The results obtained from time-dependent Cox model as given in Table 4.1 agrees with previous results in Chapter 2 which indicated that patients with higher viral load levels have lesser survival rates. Table 4.2 also confirms that viral load is significantly associated with survival process of viral load patients, and thus shared parameter model is appropriate to analyse the data. The joint model was built using JM package in R incorporating patients age at baseline, type of care facilities, clinical stages of AIDS according to World Health Organisation (WHO), CD4 cell counts, districts, and previous opportunistic infections before initiation of ARV treatment. The results is provided in Table 4.2. The joint model results

Table 4.11: Prediction of conditional probabilities for event based on 200 Monte Carlo samples

	Time	Mean	Median	Lower	Upper
1	24.0000	1.0000	1.0000	1.0000	1.0000
1	25.4353	0.9779	0.9793	0.9603	0.9884
2	28.6148	0.9301	0.9345	0.8745	0.9633
3	31.7942	0.8839	0.8908	0.7921	0.9389
4	34.9736	0.8393	0.8478	0.7139	0.9151
5	38.1530	0.7965	0.8063	0.6404	0.8918
6	41.3324	0.7554	0.7663	0.5742	0.8690
7	44.5118	0.7160	0.7276	0.5159	0.8466
8	47.6912	0.6784	0.6902	0.4628	0.8242
9	50.8706	0.6424	0.6538	0.4144	0.8006
10	54.0500	0.6081	0.6191	0.3704	0.7772
11	57.2295	0.5754	0.5857	0.3305	0.7541
12	60.4089	0.5443	0.5540	0.2943	0.7313
13	63.5883	0.5147	0.5234	0.2615	0.7092
14	66.7677	0.4866	0.4945	0.2320	0.6879
15	69.9471	0.4599	0.4674	0.2053	0.6667
16	73.1265	0.4345	0.4416	0.1813	0.6454
17	76.3059	0.4105	0.4161	0.1598	0.6251
18	79.4853	0.3877	0.3919	0.1405	0.6055
19	82.6648	0.3661	0.3688	0.1233	0.5865
20	85.8442	0.3456	0.3470	0.1079	0.5681
21	89.0236	0.3262	0.3272	0.0942	0.5503
22	92.2030	0.3078	0.3077	0.0820	0.5331
23	95.3824	0.2903	0.2890	0.0713	0.5165
24	98.5618	0.2738	0.2716	0.0632	0.4987
25	101.7412	0.2581	0.2554	0.0567	0.4794
26	104.9206	0.2433	0.2391	0.0509	0.4603
27	108.1000	0.2292	0.2229	0.0459	0.4414

agree with preliminary analysis that those patients with higher viral loads levels have lesser survival rates. However, the districts did not have statistical significance to survival rates of HIV+ infected patients in Limpopo Province. The Community Health Care Centres, Provincial Tertiary Hospital were statistically significant in survival rates of Limpopo AIDS patients as compared with patient care in clinics, whereas, districts hospitals, regional hospitals, and specialised psychiatric hospitals were not statistically significant. The age, pre-

Table 4.12: Conditional probabilities for events

	Time	Prediction Survival
1	24.0000	1.0000
1	25.4353	0.9787
2	28.6148	0.9324
3	31.7942	0.8874
4	34.9736	0.8438
5	38.1530	0.8015
6	41.3324	0.7606
7	44.5118	0.7212
8	47.6912	0.6833
9	50.8706	0.6468
10	54.0500	0.6118
11	57.2295	0.5783
12	60.4089	0.5462
13	63.5883	0.5156
14	66.7677	0.4864
15	69.9471	0.4585
16	73.1265	0.4320
17	76.3059	0.4068
18	79.4853	0.3828
19	82.6648	0.3601
20	85.8442	0.3386
21	89.0236	0.3182
22	92.2030	0.2989
23	95.3824	0.2806
24	98.5618	0.2633
25	101.7412	0.2470
26	104.9206	0.2316
27	108.1000	0.2171

Table 4.13: Prediction of conditional probabilities for event based on 200 Monte Carlo samples

	Time	Mean	Median	Lower	Upper
1	27.0	1.0000	1.0000	1.0000	1.0000
1	27.5	0.9466	0.9501	0.9042	0.9720
2	28.0	0.9392	0.9431	0.8908	0.9681

vious opportunistic infection, CD4 cell counts, and AIDS clinical stages were statistically significant, and they agree with our research results in Table 4.1.

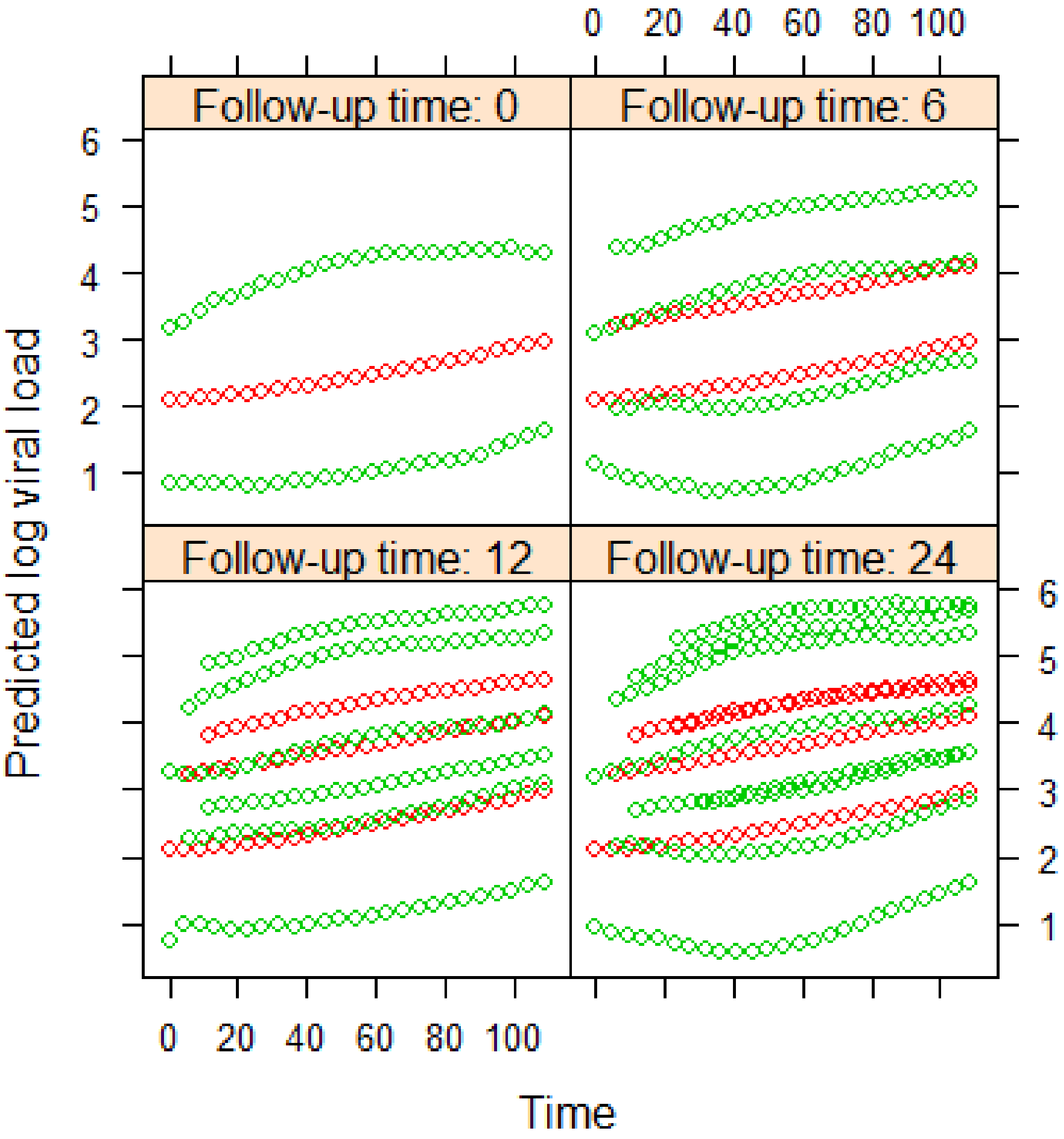


Figure 4.15: Dynamic predictions of longitudinal response for patient 18 from AIDS dataset.

The significance of the shared parameter that links the two processes (survival and longitudinal), and the reduction in the standard error of the parameters estimates when compared to independent sub-models estimates, indicates the necessity for joint model analysis of this dataset compared with the use of independent of sub-models(Cox model and linear mixed-effects model).

The viral load are a well-known biomarker of AIDS progression collected after treatment. The JLCM on viral load using Limpopo Department of Health AIDS dataset aws illustrated. JLCM has proven to be a greater asset in prediction studies since it approximates any structure, even complex, of correlated data without prior assumptions. Since big dataset is used the dynamic predictive model was more reliable and estimates of the parameters were precise and less uncertain in individual predictions, as a results, the model satisfied the validation of the dynamic predictive tools on external dataset with different characteristic that is recommended in literature [180]. Furthermore, validation of dynamic predictive tools from joint model on external data, and comparison with other approaches such as survival models including the previous biomarker measures are described in [180].

4.8 Conclusion

Whenever, the longitudinal and survival process are correlated, valid influence can be expected through the use of joint modelling approach. This has been successfully demonstrated using Limpopo Department of Health HIV/AIDS dataset to simultaneously model viral load fluctuations over time and their survival. The use of joint model compared with independent sub-models (survival and linear mixed-effects models) indeed showed a significant decrease in the standard errors. This reduction in biasness means that more accurate influence can be made using joint models parameter estimates.

Finally, all models fitted, tables, figures and parameters were estimated using available R package: lcmm for JLCM and JM package for SREM in this Chapter 4.

Chapter 5

Conclusion

5.1 Thesis summary

The thesis was organised into four chapters with a view of analysing the factors that contribute to the spread of HIV in Limpopo Province. Chapter 1 looked at the definition of HIV, transmission of HIV/AIDS, prevention of HIV infection based on literature, status of HIV in South Africa, groups that are mostly affected by HIV/AIDS in South Africa. It also, looked at factors that contribute to the spread of HIV/AIDS globally according to literature. Furthermore, the survival sub-models, linear mixed-effects sub-models and joint models in the analysis of longitudinal data was introduced. However, more focus was placed on the analysis of longitudinal and time-to-event outcomes motivated by HIV/AIDS Limpopo department of health dataset.

In Chapter 1, HIV prevention initiatives by South African National Health Department which cascades to Provincial Health Department are having a significant impact on mother-to-child transmissions, which are falling dramatically. New HIV infections overall has fallen drastically, the short term financing of South African HIV epidemic is secure, however, for longer term, the government needs to explore other strategies in order to sustain and expand its progress.

In Chapter 2, both semi-parametric (Cox proportional hazard models) and parametric (Accelerated failure time) survival models to HIV/AIDS dataset obtained from Limpopo Department of Health were fitted. Furthermore, comparing these models and it was found that generalised gamma model provided a better fit to the study of Limpopo HIV/AIDS data with lowest AIC value as a result generalised gamma was a preferred model. To explore and identify potential risk factors for the HIV-infected patient's survival rate, survival data was used.

In Chapter 3, the linear mixed-effects models to longitudinal data using viral load as biomarker was fitted. The evolution of each HIV/AIDS patient in time could be described effectively by linear mixed models. Each patient in the population has his/her patient-specific mean response profile over time. The quadratic linear mixed effects models were found to fit the data well. That the potential of curvature trajectory fit the data better. Linear mixed-effects models were used to evaluate the progression of patient's viral load. The covariates were age(at baseline), districts, AIDS clinical stages, type of health care facilities, CD4 cell counts and previous opportunisticinfections.

In Chapter 4, HIV/AIDS dataset was analysed using joint models. Joint models utilise both longitudinal and survival data. The joint model enabled to determine the strength of association between the viral load (biomarker) and risk for an event. One of the advantage of joint model over survival sub-model and linear mixed sub-model is its dynamic prediction ability for longitudinal outcome. The use of joint model compared with independent sub-models (survival and linear mixed-effects models) indeed showed a significant decrease in the standard errors, consequently, the reduction of biasness and accurate parameter estimates. More focus focused was on shared random-effects models that included characteristics of longitudinal biomarker as a predictor in the model for time-to-event. A less well-known approach of joint latent model was used, and it was assumed that a latent class structure entirely captures the correlation between the longitudinal biomarker trajectory and risk of the event using Limpopo Department of Health HIV/AIDS dataset.

All the fifteen major objectives mentioned in chapter 1 under introduction section have been addressed successfully. In Chapter 2, the use of Kaplan-Meier in comparing the average evolutions between gender, districts, health care facilities, previous opportunistic infections, and AIDS clinical stages was addressed; comparing the semi-parametric and parametric models; and, analysing survival data using both Cox proportional hazard and parametric hazard model were also addressed.

In Chapter 3 addressed the following objectives: the relationship between response variable and the covariates using linear mixed effect models and showed how longitudinal evolution of viral load is associated with time-to-death. Also, the characteristics of viral dynamics in patients population the intra- and inter-subject variation and assume random effects that gives some structure to error terms that characterises individual variation due to some factor levels was addressed. Finally, non-linear statistical framework as a basis for estimation of population and individual viral dynamics parameters was demonstrated and how these models may be used to draw biological relevant interpretations and aid clinical decision-making within the context of Limpopo HIV/AIDS dataset.

In chapter 4, the following objectives addressed: separate and joint longitudinal and survival analyses per outcome; established the strength of association between the longitudinal evolution of viral load and hazard rate to death; comparison of separate and joint models, and various association measures such as parametric joint and shared parameter models approach; compared the average evolutions between males and females; we showed how marker-specific evolutions are related to each other; computation of prediction for time to death for any randomly selected HIV positive patient by considering patient's viral load; and coming up with good joint model(s) that will handle simultaneously both the repeated measurements as well as the survival outcomes in the presence of clustering in the Limpopo HIV/AIDS dataset.

5.2 Summary of the key findings

The focus of the thesis was methodological aspect of survival, linear mixed-effect and joint models. The results are believed to better reflect the reality in a rural, peri-urban and urban of Limpopo Province, and , maybe could be applicable to other provinces of South Africa with similar setting. Stigma and delay in seeking health care, lack of voluntary testing and counsel services, and system delays in referral and ART initiation are perceived to be the major possible reasons for continued progression to advanced stages of HIV/AIDS. In our findings Mopani district stood out to be the Limpopo district with highest patients with HIV prevalence since it has more villages and mines.

Tuberculosis is found to be highly associated with the increase risk of AIDS and death in HIV-infected patients. Our findings support the view that prolonged immune activation induced by TB leads to prolonged increase of HIV replication and consequently accelerated disease progression. The TB prophylaxis drug could substantially reduce TB mortality and morbidity among those with HIV and that is particularly important in the context of copper mines in Phalaborwa, platinum mine in Mokopane, Burgerfort, Thabazimbi and Northam, where the high rate of silicosis and HIV may lead to a situation in which the incidences of TB is about 3000 per 100000 men per year [83]. More importantly, there has been few intervention programmes by Provincial Government, even on small scale, which attempt to reduce transmission among migrants and their rural or peri-urban or urban partners. Policy issues need to be addressed, including the nature and extent of migration, the rights of migrants, and the kind of services they have access to.

The Cox regression proportional hazard models and Accelerated Failure Time(AFT) models have been compared using HIV/AIDS patients data of Limpopo Province. The AFT model was fitted and diagnosed using Cox-Snell residuals, and the Weibull model provided a better fit to the studied Limpopo HIV/AIDS data with a lowest AIC value, as a result the Weibull model is a preferred model. The results obtained from Kaplan-Meier curves show that males survival time were shorter than their counterpart. Thus, females have significant longer survival times as whole as compared to their counterpart. That could be attributed to males life styles such as failing to take ARV as prescribed etc.,

The survival probability of any randomly selected HIV/AIDS patients in using dynamic prediction of joint models was predicted. The individualised prediction is quite important in the current technological era, because the physicians are able to take an informed decision about patient's disease dynamics, and, the patient information can be updated as time progresses.

5.3 Limitations of the thesis

The major limitation of joint models observed so far is that it requires large computational power, thus resulting in slow convergence rate. Secondly, data obtained from Limpopo Health Department did not contain education level, race, marital status, occupation and or ethnicity of patients. These factors are very important to be considered for health decision- and policy makers in order to plan and draft policy to combat the scourge of HIV/AIDS.

5.4 Future research directions

The future research is needed to improve the estimation techniques of survival, linear mixed-effects, and joint models. Secondly, to develop or improve the computational prediction ability on survival probabilities of patients for much longer time than presently. This information of survival probabilities of patients is vital to physicians to gain a better understanding of the disease dynamics.

Research sould be conducted to include both the formal and informal sectors including illegal migrants; illegal migrants should also be able to access the health services without fear of exposure. AIDS clinical stages as time dependent covariates should be considered for future research. Again, the association between treatment (ARV) and virology of HIV+ patients should also be considered for future research.

References

[1] AIDS.gov (2016). What is HIV/AIDS. (https://http://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids).

[2] How is HIV/AIDS transmitted? (2017). (http://sfaf.org/hiv-info/basic/how-is-hivtransmitted.html).

[3] Prevention of HIV Infection. (2014). (https://www.health24.com/Medical/HIV.../Prevention-of-HIV-infection-20140530)

[4] South African National AIDS Council. (2015). Global AIDS response Progress report. (https://sanac.org.za /wp-content/uploads2016/06/GARPR report-high-res-for-print-June-15-2016.pdf).

[5] Mother to Child transmission.(2017). (https://www.westerncape.gov.za/service/prevention-mother-child-transmission-pmtct.).

[6] UNAIDS. (2014). The GAP Report 2014. (https://www.unaids.org/sites/default/files/media asset/UNAIDS Gap report en.pdf).

[7] Slabbert, M., Venter, F., Gay, C., Roeloffseen, C., Lalla-Edward, S. and Rees, H. (2017). Sexual and reproductive health outcomes

among female sex workers in Johannesburg and Pretoria, South Africa: Recommendations for public health programmes. BMC Public HealthBMC series – open, inclusive and trusted 17 (Suppl 3) :442 https://doi.org/10.1186/s12889-017-4346-0.

[8] Duby, Z., Nkosi, B., Scheibe, A., Brown,B. and Bekker, L.G. (2018). Scared of going to the clinic:Contextualising healthcare access for men who have sex with men, female sex workers and people who use drugs in two South African cities. Southern African Journal of HIV Medicine, 19 (1), 701.https://doi.org/10.4102/sajhivmed.v19i1.701.

[9] Poteat, T., Ackerman, B., Diouf, D., Ceesay, N., Mothopeng, T., and Odette, K.Z. (2017). HIV prevalence and behavioral and psychosocial factors among transgender women and cisgender men who have sex with men in 8 African countries. PLoS Med 14(11): e1002422. https://doi.org/10.1371/journal.pmed.1002422.

[10] Scheibe, A., Makapela, D., Brown, B., dos Santos, M., Hariga, F., Virk, H., Bekker, L.G., Lyan, O., Fee, N., Molnar, M., Bocai, A., Eligh, J., and Lehtovuori, H. (2016). HIV prevalence and risk among people who inject drugs in five South African cities. International Journal of Drug Policy 30, 107–115.

[11] UNAIDS. (2017). Data Book.

[12] South African AIDS Council (SANAC) (2017). National strategic Plan. 2017-2022.

[13] HIV Statistics South Africa-Deaths, antiretroviral treatment 2010-2015. (https://www.tbfacts.org/hiv-south-africa/).

[14] Van Wyk, B. (2003). Dark side of the rainbow: the impact of HIV/AIDS on the African renaissance. Centre for the Stusdy of AIDS.

[15] Lim, H.J., Mondal, P., and Skinner, S. (2013). Joint Modelling of Longitudinal and event data: Application to HIV study.

[16] Ratcliffe, S.J., Guo, W., and Ten Have, T.R. (2004). Joint Modelling of longitudinal and survival data via a common frailty. Biometrics, 60(4), 892-899.

[17] Collett D. (2015). Modelling Survival Data in Medical Research. 3rd ed. Chapman and Hall.

[18] Klein, J.P., and Moeschberger, M.L. (1997). Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer.

[19] Liang, K.Y., Self, S.G., Bandeen-Roche, K.J. and Zeger, S.L. (1995). Some recent developments for regression analysis of multivariate failure time data. Lifetimes data analysis, 1(4), 403-415.

[20] Fleming, T.R. and Harrington, D.P. (1991). Counting processes and survival analysis, Wiley, New York.

[21] Cox, D.R (1972). Regression problems and life tables. Journal Royal Statistical Society, B , 34, 187-220.

[22] UNAIDS. (2014). The GAP Report. (https://www.unaids.org/sites/default/files/media asset/UNAIDS Gap report en.pdf').

[23] South African National AIDS Council. (2015). Global AIDS Response Progress Report. (https://sanac.org.za/wp-content/uploads2016/06/GARPR report-high-res-for-print-June-15-2016.pdf').

[24] Posel D. (2005). Sex, death and the fate of the nation: reflections on the politicisation of sexuality in post-apartheid South Africa. Journal of the international African Institute, 5 (2), 125-153.

[25] Stevenson, M. (2007). An introduction to Survival Analysis. EpiCentre, IVABS, http://epicentre.massey.ac.nz.

[26] Joint United Nations Programme on HIV/AIDS (UNAIDS). (2002). Report on the global HIV/AIDS epidemic. http://www.unaids.org.

[27] Mabunda, G. (2004). HIV knowledge and practices among rural South Africa. Journal of Nursing Scholarship. 36(4), 300-3004.

[28] Kalbfleisch, J.D., and Prentice, R.L. (2011). The Statistical Analysis of Failure Times Data. Wiley-Interscience, New York.

[29] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of American Statistical Association. 72, 557-565.

[30] Breslow, N. (1974). Covarience analysis of censored survival data. International Biometric Society, 30(1), 89-99.

[31] Verbeke, G., and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal data. New York: Springer.

[32] Duchateau, L., and Janssen, P. (2007). The Frailty Model. New York: Springer.

[33] Therneau, T.M., Grambsch, P.M., and Fleming, T.R. (1990). Martingale-based residuals for survival models. Biometrika, 77, 147-160.

[34] Borgan, O., and Liest , K.A. (1990). Notes on confidence interval and bands for the survival curves based on transformations. Scandanavian Journal of Statistics, 17, 35-41.

[35] Klein, J.P. (1991). Small-Sample Moments of some estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators. Scandinavian Journal of Statistics, 18(4), 333-340.

[36] Dageid, W. (2002). I am just like anybody. Balancing normality and social death-South African women living with HIV/AIDS. Master's thesis, Department of Psychology, University of Oslo, Norway.

[37] Lazarus, R.S. (1990). Stress, coping, and illness. (In H.S Friedman Ed., Personality and disease) (pp. 97-120). New York:Springler.

[38] Marais, H. (2002). To the edge. AIDS Review (2000). Pretoria, South Africa: University of Pretoria.

[39] Pakenham, K.I., and Rinaldis, M. (2001). The role of illness, resources, appraisal, and coping stragies in adjustment to HIV/AIDS: The direct and buffering effects. Journal of Behavioural Medecine, 24(3), 13-17.

[40] Mekonnen, Y., Sanders, E., Akliklu, M., Tsegaye, A., Rinke De Wit, T.F., Tobias, F.A., Schaap, A., Wolday, D., Geskus, R., Coutinho, R.A., and Fontanet, A,L. (2003). Evidence of change in sexual behaviours among male factory workers in Ethopia. AIDS 2003, 17, 223-331.

[41] Shisana, O., Rehle, T., Simbayi, L., Labadarios, D., Jooste, S., Davids, A., Ramlagan, S., Mbelle, M., Zuma, K., van Zyl, J., Onoya, D., and Wabiri, N. (2012). South Afri Africa National HIV prevalence, HIV incidence and behaviour survey. Cape Town: HSRC Press.

[42] Collinson, M.A., Tollman, S.M., Kahn, K., Clark, S.J., and Garenne, M. (2006). Highly prevalent circular migration: households, mobility and economic status in rural South Africa. Johannesburg: Wits University Press.

[43] Shisana, O., Zungu-Dirwayi, N., Toey, Y., Simbayi, L.C., Malik, S., and Zuma, K. (2004). Marital status and risk of HIV infection in South Africa. South African Medical Journal, 94, 537-543.

[44] Rao, V.K., Ladermarco, E.P., Fraser, V.J., and Kollef, M.H. (1998). The impact of comorbidity on mortality following in-hospital diagnosis of tuberculosis. Chest, 114, 1244-1252.

[45] Hansel, N.N., Wu, A.W., Chang, B., Diette, G.B. (2004). Quality of life in tuberculosis: Patient and provider perspectives. Quality of Life Research, 13, 639–652.

[46] Silva, D.R., Menegotto, D.M., Schulz, L.F., Gazzana, M.B., and Dalcin, P.T. (2010). Factor associated with mortality in hospitalised patients with newly diagnosed tuberculosis. Lung, 188(1), 33–41.

[47] Peter, J.G., Theron, G., Singh, A., Singh, V., and Dheda, K. (2013). Sputum induction to aid the diagnosis of smear-negative or sputum-scarce TB in adults from a HIV-endemic setting. European Respiratory Journal, 43, 185–194.

[48] Mabunda, T.E., Ramalivhana, N.J., Dambiysya, Y.M. (2014). Mortality associated with tuberculosis/ HIV co-infection among patients on TB treatment in Limpopo Province, South Africa. African Health Sciences, 14(4), 849-854.

[49] World Health Organization (WHO). (2011). Global Tuberculosis Control: http://www.who.int/tb/publications/globalreport.

[50] Nakhaee, F., and Law, M. (2011). Parametric modelling of survival following HIV and AIDS in the era of highly active antiretroviral therapy: Data from Australia. Eastern Mediteterranean Health Journal, 17, 231-237.

[51] Mageda, K., Leyna, G.H., and Mmbaga, E.J. (2012). High initial HIV/AIDS-related mortality and its predictors among patients on antretroviral therapy in the Kagera region of Tanzania: A five-year retrospective cohorts study. AIDS Research and Treatment,Volume 2012, Article ID 843598, 7 pages.

[52] Farzadegan , H., Hoover, D.R., Astemborski, J., Lyles, C.M., Margolick, J.B., Markham, R.B., Quinn, T.C., and Vlahov, D. (1998). Sex differences in HIV-1 viral load and progression to AIDS. The Lancet, 352,1510-1514.

[53] Ramafedi, G., and Lauer, T. (1995). Survival trends in adolescents with human immunodeficiency virus infection. Archives of Pediatrics & Adolescent Medicine, 149(10), 1093-1096.

[54] Patel, K., Kay, R., and Rowell, L. (2006). Comparing proportional hazards and accelerated failure time models: an application in influenza. Pharmaceutical Statistics, 5, 213-224.

[55] Kay, R. (2002). On the use of the accelerated failure time model as an alternative to the proportional hazards models in the treatment of time to event data: a case study of influenza. Drug Information Journal, 36, 571-579.

[56] Gelber, R.D., Goldhirsch, A., and Cole, B.F. (1993). Parametric extrapolation of surviavl estimates with applications to quality of life evaluation of treatments. Controll Clinical trails, 14, 485-489.

[57] Lawless, J.P. (1998). Parametric models in Survival Analysis. Encyclopedia of Biostatistics, John Wiley & Sons.

[58] Oakes, D. (1997). The asymptotic information in censored survival data. Biometrika, 64, 441-448.

[59] Nawumbeni, D.R., Luguterah, A., and Adampah, T. (2014). Performance of Cox Proportional Hazard and Accelerated time models in the analysis of HIV/TB co-infection survival data. Research of Humanities and Social Sciences. 4(21), 335-430.

[60] Ghate, M., Deshpande, S., Tripathy, S., Godbole, S., Nene, M., Thakar, M., Risbud, A., Bollinger, B., and Mehendal S. (2011). Mortality in HIV

infected individuals in Pune, India. Indian Journal Medical Research, 133, 414-420.

[61] Rai, S., Mahapatra, S., Sircar, S., Raj, P.Y., Venkatesh, S., Shaukat, M., Rewari, B.B. (2013). Adherence to antitroviral therapy and its effect on survival of HIV-infected individuals in Jharkhand, India. PLoS ONE 8(6), e66860.

[62] Kaplan, E.L., and Meier, P. (1958). Non-parametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481.

[63] Helwig, N.E. (2017). Linear mixed-effects regression, University of Minnesota (USA).users.stat.umn.edu/ helwig/notes/lmer-Notes.pdf.

[64] Kee, M.K., (2009). Improvement in survival among HIV-infected individuals in the Republic of Korea: Need for an early HIV diagnosis. BMC infectious Disease, 9, 128-128.

[65] Cooper, D.A. (2008). Life expectancy of individuals on combination antiretroviral therapy in high-income countries: A collaborative analysis of 14 cohort studies. The Lancet, 372 (9635), 266-267.

[66] Rajasekaran, S., Jeyaseelan, L., Raja, K., Vijila, S., Krithigapriya, K.A., and Kuralmozhi, R. (2009). Increase in CD4 cell counts between 2 and 3.5 years initiation of antiretroviral therapy and determinants of CD4 progression in India. Journal of Postgraduate Medicine, 55(4), 261-266.

[67] Saach, S.L. (2007). The properties of the unique age-associated B cell behaviour reveal a shift strategy of immune response with age. The Lancert, 472, 293-299.

[68] Bachani, D., Garg, R., Hegg, L., Rajasekaran, S., Desphande, A., Emmanuel, P.C., and Rao, K.S. (2010). Two-year treatment outcomes of pa-

tients enrolled in India's national first-line antiretroviral therapy programme. The National Medical Journal of India, 23(1), 7-12.

[69] Alioum, A., Leroy, V., Commenges, D., Dabis, F., Salamon, R., and Aquitaine, G. (1998). Effect of Gender, Age, Transmission Category, and Antiretroviral Therapy on the Progression of Human Immunodeficiency Virus Infection using Multistate Markov models. Epidemiology, 9(6), 605-612.

[70] Nardi, A., and Schemper, M. (2003). Comparison and parametric models in clinical studies. Statistics in Medicine, 22, 3597-3610

[71] Crowley, T.R., and Hu, M. (1977). Covariance analysis of heart transplant survival data. Journal of the American Statistical Association, 72(357), 27-36.

[72] Ibrahim, J.G., Chu, H., and Chen, L.M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. Journal of clinical oncology, 28(16), 2796-2801.

[73] Cornell, M., Schomaker, M., Garone, D.B., Giddy, J., Hoffmann, C.J., Maskew, M., Prozesky, H., Wood, R., Johnson, L.F., Egger, M., Boulle, A., and Myer, L. (2012). Gender Differences in Survival among Adult Patients Starting Antiretroviral Therapy in South Africa: A Multicentre Cohort Study. PLOS Medicine 9(9): e1001304. https://doi.org/10.1371/journal.pmed.1001304.

[74] Muula, A., Ngulube T., Siziya, S., Makupe, C.M., Umar, E., Prozesky, H.W., Wiysonge, C.S., and Mataya, R.H. (2007). Gender distribution of adult patients on highly active antiretroviral therapy(HAART) in Southern Africa: a systematic review. BioMedical Central Public Health, 7, 63.

[75] Hawkins, C., Chalamilla, G., Okuma, J., Spiegelman, D., Hertzmark, E., Aris, E., Ewald, T., Mugusi, F., Mtasiwa, D., and Fawzi, W. (2011). Gen-

der difference in antiretroviral treatment outcomes among HIV-infected adults in Dar-es-Salaam, Tanzania. Aids, 25, 1189-1197.

[76] Taylor-Smith, K., Tweya, H., Harries, A., Schoutene, E., and Jahn, A. (2010). Gender difference in retention and survival on retroviral therapy of HIV-1 infected adults in Malawi. Malawi Medical Journal, 22(2), 49-56.

[77] Ochieng-Ooko, V., Ochieng, D., Sidle, J.E., Holdsworth, M., Wools-Kaloustian, K., Siika, A.M., Yiannoutsos, C.T., Owiti, M., Kimaiyo, S., and Braitstein, P. (2010). Influence of gender on loss to follow-up in a large HIV treatment programme in Western Kenya. Bull World Organ, 88, 681-688.

[78] United Millenennium Declaration, General Assembly Resolution 55/2, September 8, (2000). (www.ohcr.org/english/law/millennium.htm).

[79] Caldwell, J., and McDonald, P. (1982). Influence of Marternal Education on infant and child mortality: Levels and causes. Health Policy and Education, 2(3-4), 251-267.

[80] Ayalew, B. (2017). Mortality and its predictors among HIV infected patients taking Anti-retroviral Treatment in Ethopia: A Systematic Review. AIDS Research and Treatment, 2017, Article ID 5415298, 10 pages.

[81] Damtew, B. (2015). Survival and determinants of mortality in adult HIV/AIDS patients initiating antiretroviral therapy in Somalia Region, Eastern Ethopia. Pan African Medical Journal, 22, 138.

[82] Shabangu, P., Beke, A., Manda, S., and Mthethwa, N. (2017). Predictors of survival among HIV-positive children on ART in Swaziland. African Journal of AIDS Research, 16(4) , 335-343.

[83] Corbett, E.L., Charalambous, S., Moloi, V.M., Fielding, K., Grant, A.D., Dye, C., De Cock, K.M., Hayes, R.J., Williams, B.G., and Churchyard,

G.J. (2004). Human Immunodeficiency virus and the prevalence of undiagnosed tubercolosis in African gold Miners. American Journal for respiratory and clinical care medicine, 170(6) Open access.

[84] Klausner, J.D., Serenata, C., O'Bra, H., Mattson, C.L., Brown, J.W., Wilson, M., Mbengashe, T., and Goldman, T.M. (2011). Scale-up and continuation of ART therapy in South African treatment programmes, 2005-2009. Journal of acquired immune defieciency syndrome. 56(3), 292-295.

[85] Shisana, O., Hall, E.J., Maluleke, R., Chauveau, J., and Schwabe, C. (2004). HIV/AIDS prevalence among South African health workers. South African Medical Journal, 94, 846-850.

[86] Laird, N.M., and Ware, J.H. (1982). Random effects models for longitudinal data. Biometrics, 38, 963-964.

[87] Wang, W., and Heckman, N. (2009). Identifiable in linear mixed models. Technical report, Department of Statistics, University of the British Columbia.

[88] Jennrich, R.I., and Schluchter, M. (1986). Unbalanced repeated-measures models with structured covarince matrices. Biometrics, 42, 805-820.

[89] Lindstrom, M.J., and Bates, D.M. (1990). Non-linear mixed effects models for repeated measures data. Biometrics, 46, 673-687.

[90] Lindstrom, M.J., and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Assosiation, 83, 1014-1022.

[91] Singer, J.D., and Willett, J.B. (2003). Applied Longitudinal Data Analysis. Oxford Press, New York.

[92] Aiken, L.S., and West, S.G. (1991). Multiple Regression: Testing and Interpreting Interactions. Sage, Newbury Park, CA.

[93] Cillessen, A.H.N., and Borch, C. (2006). Developmental trajectories of adolescent popularity: a growth curve modeling analysis. Journal Adolescent, 29, 935-959.

[94] Singer, J.D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics, 23(4), 323-355.

[95] LeMay, V.M. (1990). A Linear least technique for fitting a simultaneous system of equations with a generalised error structure. Canadian journal of forest research, 20, 1830-1839.

[96] Steel, R.G.D., Torrie, J.H., and Dickey, D.A. (1997). Principle and Procedures of Statistics: A Biometric Approach. McGraw-Hill, New York.

[97] Gregoire, T.G., Schabenbergere, O., and Barrett, J.P. (1995). Linear modeling of irregularly spaced, unbalanced, longitudinal data from permanent-plot measurements. Canadian journal of forest research, 25, 137-156.

[98] Fortin, M., and Ung, C.H., Begin, J., and Archambault, L. (2007). Variance-covariance structure to take into account repeated measurements and hetroscedasticty in growth modeling. European Journal Forest Research, 126, 573-585.

[99] Pinheeiro, J.C. and Bates, D.M. (2000). Mixed-Effects Models in S and S-Plus. Springer, Heildelberg.

[100] West, B., Welch, K., and Galeck, A. (2007). Linear Mixed models: A Practical Guide Using Statistical Software. Chapman and Hall, Boca raton, FL.

[101] Wittekind, A., Rader, S., and Grote, G. (2010). A longitudinal study of determinants of perceived employability. Journal of Organizational Behavior, 31(4), 566-586.

[102] Wolfinger, R.D. (1996). Heterogeneous variance-covariance structures for repeated measures. Journal of Agricultural, Biological, and Environmental Statistics, 1(2), 205-230.

[103] Kigozi, B.K., Sumba, S., and Mudyope, P. (2009). The effect of AIDS defining conditions on immunological recovery among patients initiating antiretroviral therapy at Joint Clinical Research Centre, Uganda. AIDS Research and Therapy. 6, article 17, Open Access.

[104] Amuron, B., Levin, J., and Birunghi, J. (2011). Mortality in an antiretroviral therapy programme in Jinja, South-East Uganda: a prospective cohort study, Uganda. AIDS Research and Therapy. 6, article 17, Open Access.

[105] Suchindran, S., Brouwer, E.S., Van Rie, A. (2009). Is HIV infection a risk factor for multi-drug resistant tubercolosis? A systematic review. PloS ONE, 4, 5 Open Access.

[106] Francis, D.J., Fletcher, J.M., Stuebing, K.K., Davidson, K.C., and Thompson, N.M. (1991). Analysis of Change: Modeling individual growth. Journal of Consulting and Clinical Psychology, 59(1), 27-37.

[107] Hox, J.J. (2002). Multilevel Analysis: Techniques and applications. Erlbaum, Hillsdale, NJ.

[108] Barcikowski, R. (1981). Statistical power with group mean as unit of analysis. Oxford Press, New York.

[109] Graves, S., Jr. and Frohwerk, A. (2009). Multilevel modeling and school psychology: review and practical example. School Psychology Quarterly, 24(2), 84-94.

[110] Willett, J.B. (1998). Questions and Answers in the measurement of change. Review of Research in Education, 15, 345-422.

[111] Daniel, T.L.S., and Cecilia, M.S.M. (2011). Longitudinal data analyses using linear Mixed models in SPSS: Concepts, Procedures and illustrations. The Scientif World Journal, 11, 42-76.

[112] Miner, J.L., Clarke-S., and Sterwart, A.C. (2008). Trajectories of externalising behaviour from age 2 to 9: relations with gender, temparement, ethnicity, parenting, and rater. Development Psychology, 44, 771-786.

[113] Speer, D.C., and Greenbaum, P.E. (1995). Five methods for computing significant individual client change and rates: support for an individual growth curve approach. Journal of Consulting and Clinical Psychology, 63, 1044-1048.

[114] Bryk, A.S., and Raudenbush, S.W. (1992). Hierachical Linear Models. Sage, Newbury Park, CA.

[115] Bono, R., Arnau, J., and Ballueka, N. (2007). Using linear mixed models in longitudinal studies: appplication SAS PROC MIXED. REMA Revista electrónica de metodología aplicada, 12(2), 15-31.

[116] Helwig, N.E. (2017). Linear Mixed- Effects Regression. University of Minnesota.

[117] Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004). Applied Longitudinal Analysis. A John Wiley and Sons, ING., Publications.

[118] Mabunda, T.E., Ramalivhana, N., and Dambisya, Y. (2014). Mortality with tuberculosis /HIV co-infection among patients on TB treatment in the Limpopo Province, South Africa. Africa Health Sciences, 14(4), 849-854.

[119] Badri, M., Ehrlich, R., Wood, R., Pulewitz, T. and Maartens, G. (2001). Association between tuberculosis and HIV disease progressionin a high

prevalence area. The International Journal of Tuberculosis and Lung Disease, 5(3), 225-232.

[120] Limpopo Provincial AIDS Council (LPAC). (2017): Provincial Stragic Plan 2012-2016.

[121] Fox, J. (2016). Applied Regression Analysis and Generalised linear models. Sage, Thousand Oakes, CA, thirdedition.

[122] Bates, D., Maechler, M., Balker, B., and Walker, S. (2014). Lme4: Linear Mixed-effects models using Eigen and S4: R packgage version 1, 1-7.

[123] Ragnarsson, A., Onya, H.E., Thorson, A., Ekstr$\ddot{o}$m, A.M., and Aar∅, L.E. (2008). Young males' gendered sexuality in the era of HIV and AIDS in Limpopo Province, South Africa. Qualitative Health Research, 18(6), 739-746.

[124] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the Armerican Satatistics Association, 72, 320-340.

[125] Peng, H., and Lu, Y. (2012). Model selection in linear mixed effects models. Journal of Multivariate Analysis, 109, 109-120.

[126] Hedeker, D., and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. Biometrics, 50, 933-944.

[127] Saikia, R., and Barman, M.P. (2016). Comparing Cox Proportional hazard model and parametric counterparts in the analysis of esophagus cancer patients data. IOSR Journal of Mathematics, 12(5), 16-21.

[128] Teshnizi, S.H., and Ayatollahi, S.M.T. (2017). Comparison of Cox regression and parametric models: application for assessment of survival of pediastric cases of acute leukemia in Southern Iran. Asian Pacific Journal of cancer prevention, 18(4), 981-985.

[129] Adelian, R., Jamali, J., Zare, N., Ayatollahi., S.M.T., Pooladfar, G.R., and Oustaei, N. (2015). Comparison of Cox's regression model and parametric models in evaluating the prognostic factors for survival after liver transplatation in Shiraz during 2000-2012. International Journal of organ transplatation Medicine, 6(3), 119-125.

[130] Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalised linear models. Biometrika, 73, 13-22.

[131] Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance components models. Journal of the American Statistical Association, 76, 351-353.

[132] Box, G.E.P., and Tiao, G.C. (1992). Bayesian Inference in Satatistical Analysis. Wiley Classic Library edition. John and Sons, New York.

[133] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian data Analysis, Texts in Statistical Science, Chapman and Hall,London.

[134] Laird, N.M., and Ware, J.H. (1982). Random effects models for longitudinal data. Biometrics, 38, 963-964.

[135] DeGruttolla, V., and Tu, X.M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics, 50, 1003-1014.

[136] Faucett, C.J., and Thomas, C.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates. A Gibbs sampling approach. Statistics in medicine, 15, 1663-1685.

[137] LaValley, M.P., and DeGruttola, V. (1996). Models for empirical Bayes estimators of longitudinal CD4 counts, Statistics in medicine, 15, 2289-2305.

[138] Pawitan, Y., Self, S. (1993). Modeling disease marker processes in AIDS. Journal of the American Statistical Association, 83, 719-726.

[139] Taylor, J.M.G., Cumberland, W.G., and Sy, J.P. (1994). A stochastic model for analysis of longitudinal AIDS data. Journal of the American Statistical Association, 89, 727-736.

[140] Wulfsohn, M.S., and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. Biometrics, 53, 330-339.

[141] Guo, X., and Carlin, B. (2004). Seperate and joint modelling of longitudinal and event time data using standard computer packages. The American Statistician, 58, 16-24.

[142] Chi, Y.Y., and Ibrahim, J.G. (2006). Joint models for multivaraite longitudinal and multivaraiate survival data, Biometrics, 62, 432-445.

[143] Tsiatis, A.A., and Davidian, M. (2004). Joint modelling of longitudinal and time-to-event data: an overview, Statistica Sinica, 14, 809-834.

[144] Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal maesurements and event time data. Biostatistics, 1, 465-480.

[145] Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data. Statistics in medicine, 11, 1861-1870.

[146] Xu, J., and Zeger, S.L. (2001). Joint analysis of longitudinal data comprimising repeated measures and times to events. Journal of the Royal Statistical Society, 50(3), 375-387.

[147] Song, X., Davidian, M., and Tsiatis, A.A. (2002). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. Biostatistics, 3, 511-528.

[148] Yu, M., Law, N.J., Taylor, J.M.G., and Sander, H.M. (2004). Joint longitudinal-Survival-cure models and their application to prostrate cancer. Statistica Sinica, 14(3), 835-862.

[149] Chen, M.H., Ibrahim, J.G., and Sinha, D. (2004). A new joint model for longitudinal and survival data with cure fraction. Journal of Multivariate Analysis, 91, 18-34.

[150] Self, S., and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In: Jewell N.P., Dietz K.,Farewell V.T. (eds) AIDS Epidemiology. Birkhäuser, Boston, MA.

[151] Nowak, M.A., and Bangham, C.R.M. (1996). Population dyamamics of immune response to persistent viruses. Science, 272,74-79.

[152] Perelson, A.S., Kirschner, D.E., and De Boer, R.D. (1993). Dynamics of HIV infection of CD4 T cells. Mathematical Biosciences, 114, 81-125.

[153] Phillips, A.N. (1996). Reduction of HIV concentration during acute infection: Independence from a specific immune response. Science, 271, 497-499.

[154] Schenzle, D. (1994). A model for AIDS pathogensis. Statistics in Medicine 13, 2067-2079.

[155] Tan, W.Y., and Wu, H. (1998). Stochastic model of the dynamics of CD4 T cells infection by HIV and some Monte Carlo studies. Mathematical Biosciences, 147, 173-205.

[156] Wu, H. and Ding, A.A. (1999). Population HIV-1 dynamics in Vivo: Applacable models and inferential tools for virological data from AIDS clinical trials. Biometrics, 55, 4010-418.

[157] Ho, D.D., Neumann, A.U., Perelson, A.S., Chen, W., Leonard, J.M., and Markowitz, M. (1995). Rapid turnover of plasma ririons and CD4 lymphocytes in HIV-1 infection. Nature 373, 123-126.

[158] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. Statitstics in medicine, 8, 432-440.

[159] Ibrahim, J.G., Chu, H., and Chen, L.M. (2010). Basic Concept and Methods for joint Models of Longitudinal and Survival data. Journal of Clinical Oncology, 28(16), 2796-2801.

[160] Crowther, M.J., Abrams, K.R., and Lambert, P.C. (2012). Flexible parametric joint modelling of longitudinal and survival data. Statistics in Medicine. Wiley, Library.

[161] Zhou, M. (1991). Some properties of Kaplan-Meier estimator for independent non-identically distributed random varaibles. The Annals Statistics 19(4), 2266-2274.

[162] Dempster, A., Laird, N., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

[163] Wang, Y., and Taylor, J. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. Journal of the American Statistical Association, 96, 895-905.

[164] Verbeke, G., and Lesaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed effects models for longitudinal data. Computational Statistics and Data Analysis, 23, 541-556.

[165] Tao, H., Palta, M., Yandell, B. S., and Newton, M. A. (1999). An estimation method for the semiparametric mixed effects model. Biometrics, 55, 102-110.

[166] Heagerty, P. J., and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika, 88, 973-985.

[167] Greenwood, M. (1926). The natural duration of cancer. Reports of public Health and realted subjects 33, HMSO, London.

[168] Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics, 57, 795-802.

[169] Brown, E. and Ibrahim, J. (2003). A Bayesian semi-parametric joint hierarchical model for longitudinal and survivala data. Biometrics, 59, 221-228.

[170] Rizopoulos, D., and Ghosh, P. (2011). A Bayesian semi-parametric multivariate joint model for multiple longitudinal outcomes and time-to-event. Satistics in Medicine, 30, 1366-1380.

[171] Therbneau, T., and Grambsch, P. (2000). Modeling Survival Data: Extending the Cox Model. Spring-Verlag, New York.

[172] Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data with applications in R. Chapman and Hall/CRC Biostatistics Series, Boca Raton.

[173] Andrinopoulou, E.R. (2014). Joint Modelling of Longitudinal and Survival Data with Appplication in Heart valve Data. PhD Thesis, Erasmus University.

[174] Norbre, J., and Singer, J. (2007). Residuals analysis for linear mixed models. Biometrical Journal, 6, 863-875.

[175] Verbeke, G., and Molenberghs, G. (2000). Linear mixed models for longitudinal Data. Springer-Verlang, New York.

[176] Barlow, W., and Prentice, R. (1998). Residuals for relative risk regression. Biometrika, 75, 65-74.

[177] Therneau, T., Grambsch, P., and Fleming, T. (1990). Martingale based residuals for survival models. Biometrika. 77, 147-160.

[178] Cox, D.R., and Snell, E. (1998). A general definition of residuals. Journal of the Royal Statistical Society, Series B, 30, 248-275.

[179] Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics, 67, 819-829.

[180] Proust-Lima, C., and Taylor, J. (2009). Development and validation of a dynamic pronostic tool for prostrate cancer recurrence using repeated measures of post-treatment PSA: A joint modeling approach. Biostatistics 10, 535-549.

[181] Yu, M., Taylor, J., and Sandler, H. (2008). Individualised prediction in prostrate cancer using joint longitudinal survival-cure model. Journal of American Statistical Association, 103, 178-187.

[182] Garree, F., Zwinderman, A., Geskus, R., and Sijpkens, Y. (2008). A joint latent class change-point model to improve the prediction of time to graft failure. Journal of the Royal Statistical Society, Series A, 171, 299-308.

[183] Dobson, A., and Henderson, R. (2003). Diagnostic for joint longitudinal and dropout time modeling. Biometrics 59, 741-751.

[184] Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. Biometrics, 66 , 20-29.

[185] Dempster, A., Rubin, D., and Tsutakama, R. (1981). Estimation in covariance components models. Journal of American Statistical Association, 76, 341-353.

[186] Proust-Lima, C., Sene, M., Taylor, M.G., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: Review. Statistical Methods in Medical Research, 23(1) , 74-90.

[187] Lin, H., Turnbull, B.W., McCulloch, C.E., and Slate, E.M. (2002). Latent class models for joint analysis of lonitudinal biomaker and event process data: application to longitudinal prostate-specific antigen reading and prostate cancer. Journal American Statatistical Assoc, 97, 53-65.

[188] Rubin, D. (1976). Inference and missing data. Biometrika, 63, 581-592.

[189] Little, R., and Rubin, D.B. (2002). Statistical Analysis with missing data, 2nd edition. Wiley, New York.

[190] Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method, 61 (3), 439-447.

[191] Liang, K., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. Biomatrika , 73(1), 13-22.

[192] Hartigan, J.A. (1969). Linear Bayesian methods. Journal Royal Statistcal Society, Series, B, 31, 440-454.

[193] Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized linear models. Journal Royal Statistiscal Society, A, 135, 370-84.

[194] Jorgensen, B. (1983). Maximum Likelihood Estimation and Large-Sample Inference for Generalized Linear and Nonlinear Regression Models. Biometrika, 70(1), 19-28.

[195] Morton, R. (1981). Efficiency of Estimating Equations and the Use of Pivots. Biometrika, 68(1) , 227-233.

[196] UNAIDS (2017). Ending AIDS: Progress towards 90-90-90target.